УДК 004:34 + 930.25:004

Джони Кадавид, Джонхан С. Баша, Гандхимани Калисваран

Юридические и технические проблемы сетевого архивирования в Сингапуре

Доклад на сессии «Политика формирования цифровых коллекций в национальных библиотеках: проблемы и возможности» Всемирного конгресса ИФЛА (17–23 авг. 2013 г., Сингапур).

Рассмотрены юридические и технические сложности сетевого архивирования, препятствия на пути сбора и управления электронными коллекциями. Проанализированы новейшие международные достижения и проблемы в этой сфере. Выводы и рекомендации сформулированы, исходя из функций Национальной библиотеки, которая обязана поддерживать научные исследования и накопление знаний, в том числе посредством проектных решений, заложенных в сетевом архиве.

Опыт Сингапура может быть интересен и полезен и другим странам. (Оригинал доклада на англ. яз., справочный материал и библиографию см.: http://library.ifla.org/217/)

Ключевые слова: сетевой архив, WAS, авторское право, обязательное депонирование, Сингапур, облачные технологии, сетевые вычисления, эволюция терминологии.

Сетевое архивирование возникло сравнительно недавно – в конце XX в.; его цель – избежать появления «чёрных дыр» в мировой истории. Постоянно растущее количество изначально цифровых публикаций в Интернете, сопряжённое с международным признанием того, что эти публикации являются частью культурного наследия («цифровое наследие»), побудило правительство признать обеспечение сохранности сетевых публикаций существенной проблемой национальной важности.

На Азиатском континенте только национальные библиотеки Южной Кореи, Японии, Китая и Сингапура осуществляют сетевое архивирование национальных доменов. Сложности сетевого архивирования ставят новые проблемы в управлении цифровыми коллекциями, особенно в отношении законодательства об обязательном экземпляре и авторского права. Это оказывает влияние на сбор материала в Сети и использование документов сетевого архива.

Политические и этические аспекты сетевого архивирования

Сетевое архивирование ставит этические вопросы на любом из этапов жизненного цикла работы с информацией. Что следует сохранять? Кто отвечает за сохранность? Кто решает, кому предоставлять доступ, при каких обстоятельствах и к каким документам? Подобные вопросы разрешаются в ходе работы с архивами, но существуют определённые характеристики электронных документов и сетевого архивирования, которые ставят специфические вопросы этики.

Существует большая разница между печатными и электронными публикациями в смысле их постоянства. Это – очень важный момент для понимания культурного наследия: суть проблемы сбора материалов – сохранить культурную идентичность и создать ресурсы – коллекции по науке и культуре, которые позволяют обогащать культуру, совершенствовать исследования и обеспечить широкие социальные и экономические выгоды тому обществу, которое поддерживает и финансирует сбор.

Эфемерные печатные публикации архивируются в очень небольшом количестве, но при архивировании Сети огромная доля цифровых публикаций, по характеру своему являющихся эфемеридами, может собираться и долговременно сохраняться. Это, например, сообщения Твиттера, комментарии в форумах, мнения в статьях, посты в Facebook и другие документы, созданные пользователями, причём, создатели документов не предполагали их сохранение в институтах памяти на многие поколения.

Ещё одна проблема, встающая по мере развития сетевого архивирования – это цензура, которая

контролирует интеллектуальную продукцию в Сети. Цензура может выражаться в ограничении доступа к материалам; сознательное пренебрежение веб-сайтами или нежелание их сохранять по политическим или другим причинам по сути является одной из форм цензуры. Авторское право может использоваться и по политическим мотивам с тем, чтобы заставить убрать или уничтожить какие-либо архивные материалы.

Национальная идентичность
и Национальная библиотека Сингапура

В последние годы развитие сервисов Национальной библиотеки Сингапура (НБС) происходило в соответствии с формированием общества, члены которого занимаются непрерывным обучением в течение всей жизни, что способствует созданию интеллектуального капитала и нового цикла национальных инноваций.

Наблюдательный совет НБС при формировании сетевого архива поставил перед библиотечным коллективом задачу создать электронную коллекцию материалов, произведённых в Сингапуре и опубликованных через Интернет, в которой отражались бы различные грани культурного наследия страны. Предполагается, что сетевой архив поможет: «Добиться у сингапурцев ощущения общности, национальной идентичности и ощущения своих национальных корней. Этого можно достичь, архивируя информацию, которая формирует национальную идентичность. Использование Сети в качестве инструмента социальных коммуникаций постоянно растёт. Со временем это обеспечит запись событий, касающихся среды, в которой живёт наш народ, путей его развития и эволюции самоидентификации. Архивированные записи событий окажутся бесценным источником документального наследия для нынешних и будущих поколений сингапурцев. Понимание этого формирует чувство общности и соучастия, которое укрепляется с помощью хорошо и глубоко продуманного архива» (*National Library Board, 2012 Singapore Country Report for the CDNL-AO Meeting 2012*).

Сетевое архивирование в Сингапуре

С созданием в 1995 г. Наблюдательного совета Национальной библиотеки Сингапура, ответственность за сбор обязательного экземпляра была возложена на НБС; в её функции входило комплектование и поддержка исчерпывающей коллекции библиотечных материалов, относящихся к Сингапуру и его жителям. При этом понятие «библиотечные материалы» было сформулировано в широком смысле – «любые материалы или данные, которые можно воспроизвести». Интерпретация этого определения позволяет НБС собирать в том числе и электронные документы, если они зафиксированы на физических носителях (например CD, DVD и т.п.), но сетевая информация под это определение не подпадает.

В 2004 г. Наблюдательный совет создал рабочую группу, в задачи которой входило уточнение категорий обязательного экземпляра. Группа изучила мировой опыт и рекомендовала незамедлительно начать сбор электронных документов на добровольных началах, а также построение новой структуры депонирования, и внести соответствующие поправки в Закон о Национальной библиотеке.

Формирование системы сетевого архивирования в НБС. В 2005 г. сингапурский Центр исследований Интернета (*Singapore Internet Research Center*) совместно с *Internet Archive*, и *WebArchivist.org.* инициировал создание Азиатского сетевого архива по цунами (*Asian Tsunami Web Archive*). Некоторые члены Рабочей группы, в особенности учёные из Технологического университета Наньянг в Сингапуре, приняли участие в этом проекте. Работа продлилась около двух месяцев: за это время было собрано 1 600 веб-сайтов из 40 стран (на 13 языках).

В 2006 г. на базе опыта, приобретённого в ходе работы над проектом Азиатского сетевого архива по цунами, и основываясь на рекомендации Рабочей группы, НБС начала проект Сетевого архива Сингапура *(Web Archive Singapore – WAS*) – сбор веб-сайтов национальной значимости, которые могут быть полезны будущим поколениям. Была принята селективная модель, в рамках которой отобрано около 100 сайтов – в основном правительственных учреждений и официальных организаций. В 2007 г. приступили к селективному сбору сайтов домена .sg, и к 2009 г. архивировалось уже более 500 сайтов в месяц; всего в домене зарегистрировано около 100 тыс. сайтов, из них 20 тыс. были заархивированы к 2010 г.

В 2010 г. НБС инициировала ряд проектов, облегчающих доступ к электронным ресурсам: Национальная электронная библиотека (в проекте предусмотрена оцифровка печатного фонда), выпуск энциклопедии

Singapore Infopedia, газетный проект *NewspaperSG* и проект *Singapore Memory*.

Проблемы авторского права в проекте сетевого архива. Среди международных соглашений по интеллектуальной собственности и договоров, подписанных Сингапуром, – Бернская конвенция и двусторонние соглашения о свободной торговле с США. Это создаёт определённые рамки для законодательства об авторском праве в Сингапуре. Тем не менее недавно НБС заявила, что Сингапур поддерживает гибкий подход к авторским правам:

«НБС тесно сотрудничает с Международной федерацией библиотечных ассоциаций и учреждений, Всемирной организацией интеллектуальной собственности и Офисом по интеллектуальной собственности Сингапура, лоббируя изменения в законодательстве по авторскому праву с тем, чтобы убедить законодателей в необходимости создания системы ограничений и исключений для библиотек в законодательстве по правам об интеллектуальной собственности как на международном уровне, так и в Сингапуре. НБС пересматривает действующий закон о библиотеке, чтобы включить в него положения о депонировании электронных документов» (*National Library Board*).

Не вполне ясно, сколь удачно и каким образом наши коллеги лоббируют в международных организациях, но вот на национальном уровне поправки в Закон об обязательном депонировании ожидают своего решения уже около 10 лет.

В Законе Сингапура об авторском праве сказано, что любое воспроизведение защищённой работы является нарушением, если нет разрешения от правообладателя. В Законе нет специальных исключений или ограничений в отношении сетевого архивирования или обеспечения сохранности электронных документов; нет юридической определённости и в защите сетевого архивирования в рамках концепции честного пользования.

В разделе 35 Закона честное пользование описано в общих словах, допускающих гибкие критерии для оценки допустимости нарушения закона с учётом следующих факторов: цель и характер использования, в том числе коммерческого или некоммерческого – в образовательных целях; природа произведения либо его адаптация; объём и значимость скопированной части для всего произведения или его адаптации; эффект использования на потенциальном рынке или его влияние на стоимость произведения либо адаптации; возможность получения работы или её адаптации за разумное время и по обычной коммерческой цене.

Вопросы авторского права. Проект WAS прошёл несколько этапов развития, начиная от очень узкого селективного выбора веб-сайтов и до более широкого подхода – фиксирования национального домена целиком или частично.

Фиксация веб-сайтов влечёт за собой воспроизведение их содержания. НБС в ходе селекции и просмотра сайтов придерживается политики «разрешено по умолчанию» (*opt-out policy*), делая моментальные снимки сайтов и посылая администраторам «Извещение об архивировании сайтов», в котором оговорены некоторые преимущества сетевого архивирования: «Если ваш веб-сайт и сетевые публикации будут заархивированы библиотекой, вы получите доступ к ранним версиям вашего сайта, хранящимся в сетевом архиве. Библиотека также проведёт каталогизацию ваших публикаций, тем самым делая ваш сайт более заметным в Сети, в том числе и для учёных, которые пользуются сетью нашей библиотеки». В извещении также сказано: «НБС сохраняет право изъять любой материал из сетевого архива, если материал, по её мнению, нарушает авторское право».

При архивировании материала никакой оценки чистоты в отношении копирайта не предпринимается; позднее в рамках методики «Извещения и изъятия» материал может быть удалён из архива (если кто-либо этого потребует или библиотека примет соответствующее решение). Аналогичная политика реализуется в Нидерландах.

Аналогичная технология используется в *Internet Archive*, а также предлагается в качестве предпочтительной в США. Однако существует риск судебного преследования ввиду юридической неопределённости – риск, который можно было бы избежать, приняв адекватное законодательство об обязательном экземпляре, подобное принятому в Новой Зеландии, Великобритании, Франции.

Важно подчеркнуть, что законодательство о депонировании не может быть абсолютным решением,

поскольку оно ограничено территориальными и технологическими рамками. Например, в Новой Зеландии Национальная библиотека обязана испрашивать разрешение при сборе зарубежных или местных документов, защищённых паролями.

Из архивирования исключаются материалы, не имеющие авторизации (в смысле авторских прав): «Отметим также, что изображения и документы, права на которые принадлежат третьей стороне, не могут включаться в архив из-за проблем с копирайтом» и «НБС не комплектует и не архивирует информацию, размещённую на других доменах или серверах».

Сохранение и сохранность. Как уже говорилось, в проекте принята технология изъятия по требованию правообладателя, установлены процедуры извещения и расследования, в которых рассматриваются основания для признания требования оправданным. Без законодательного подкрепления нет возможности гарантировать постоянную сохранность собранного материала, и многое зависит от соглашения с правообладателем.

Сетевой архив НБС считается частью культурного наследия страны и должен оберегаться в рамках долговременной стратегии. Это означает неоднократное использование материалов, что в принципе ведёт к нарушению действующего авторского права. Если мы хотим достигнуть поставленных целей, нужно внести соответствующие поправки в законодательство, как это сделано в США и Великобритании. Положения авторского права в Сингапуре сильно устарели и не соответствуют цифровой эпохе. НБС имеет право воспроизводить материал в целях обеспечения его сохранности, но этого недостаточно. Долговременная сохранность предполагает возможность трансформации материала с учётом новых протоколов и компьютерных платформ, что сейчас запрещено.

Использование сетевого архива. Сбор и обеспечение сохранности материалов в сетевом архиве ставят сложные вопросы, но не менее сложна проблема обеспечения доступа. Сегодня мы предоставляем доступ лишь к очень малой части коллекций: с веб-сайта WAS доступны только те материалы, на которые получено специальное разрешение, или материалы, созданные правительственными органами. Мы ждём внесения поправки в Закон о Национальной библиотеке, которая позволит размещать в Интернете большое количество материалов.

Единственная национальная библиотека, которая предоставляет открытый доступ к сетевым архивам – это Национальная и университетская библиотека Исландии (*National and University Library of Iceland;* www.Vefsafn.is), где ограничен доступ только к платным материалам. Основанием для ограничений доступа являются опасения правообладателей в том, что будет нанесён экономический ущерб их интересам.

Национальная библиотека Франции ограничивает использование материалов французского сетевого архива с применением технологии глубокого анализа данных (*data mining analysis*). Сетевые архивы используются во многих исследованиях: анализ социальной активности, историография сети, этическое воздействие архивирования и т.п. Эффективность использования зависит от полноты архива, удобства пользования и публичности. Не всегда этого удаётся достичь.

С самого начала организаторы проекта WAS и руководство НБС не сумели донести до граждан информацию о проекте. Один из жителей Сингапура, получивший приглашение передать свой сайт в архив, сказал: «Мы несколько раз звонили в библиотеку по вопросу наследия, но никто ничего не знает, меня отсылали от одного отдела к другому».

Сиротские произведения. Сиротские произведения – это те, правообладателей которых невозможно либо идентифицировать, либо определить их местонахождение. Законодательство об авторском праве устанавливает, что любое пользование защищённым произведением требует разрешения правообладателя, и исключения из закона должны быть специально оговорены. Пока что в Сингапуре не найдено решение этой проблемы, и обсуждение её не ведётся. Для сетевых архивов это означает невозможность их использования в будущем, значит, нужна либо чёткая идентификация авторства, либо соответствующие поправки в законы.

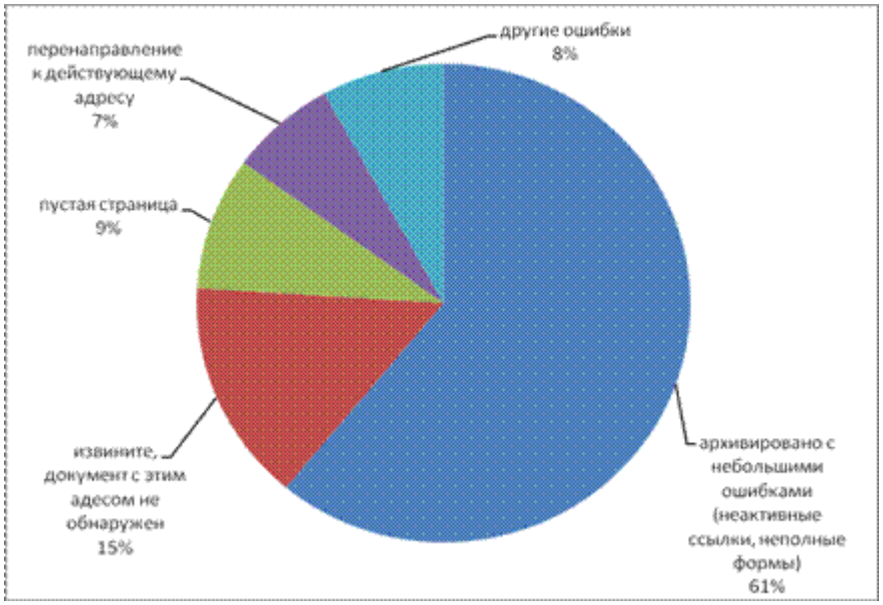
Таксономия сетевого архива

В сетевом архиве Сингапура – 1174 веб-сайта; они распределены по 11 широким категориям. Наибольшая группа – это «Организации», за ними следуют «Политики и правительство» и «Искусство». В категории

«Персоналии» – только 4 веб-сайта: большинство сайтов хорошо известных сингапурцев пока что не учитываются – нам только ещё предстоит собирать веб-сайты граждан Сингапура более полно. Не разработан и критерий, по которому ведётся отбор конкретного документа в состав коллекции культурного наследия.

http://ellib.gpntb.ru/subscribe/ntb/2014/9

На примере типовых ошибок в категории «Искусство» (см. рис.) показано невысокое качество содержания архива WAS, следовательно, необходимо внести усовершенствования в процесс архивирования.



Наиболее типичные ошибки архивирования в категории «Искусство»

Поиск и прикладные программы. Потребности и интересы опытных пользователей сильно различаются. Преподаватели и учёные могут искать сетевые документы для проведения исследований – они будут ссылаться на эти документы или цитировать их в своих трудах. Для широкой публики, использующей Интернет в качестве коммуникативной среды, сетевой архив, быть может, не столь и важен. В настоящее время большинство документов архива сохраняется для научных целей.

В любом случае пользователь надеется, что архив обладает теми же поисковыми возможностями, что и «живая» Сеть, поэтому существует настоятельная потребность разработки надёжных способов выхода на архивные материалы. Большинство интерфейсов сетевых архивов пока что ограничиваются системами на базе URL, причём основные поисковые функции требуют ручного набора адресов URL и категорий. Такие поисковые машины, как *Wayback Machine*, *WERA* (*Web Archive Access*) и *Hanzo-WARC* обеспечивают лишь базовые функции.

На современном уровне развития технологий сетевого архивирования удаётся индексировать только текстовые документы; аудио- видеофайлы и файлы изображений не индексируются. Индексирование текста происходит автоматизированным путём на основе метаданных в заголовках, сгенерированных кроулером. В некоторых интерфейсах сетевых архивов для индексирования и поиска используются возможности системы *Google*.

Поисковые машины обеспечивают автоматическое индексирование – это очень важно, поскольку удаётся управлять поиском и проводить его в миллиардах документов. Однако для более полного удовлетворения поисковых потребностей сетевых архивов требуются специальные машины (например, *NutchWax* – усовершенствованная версия поисковика *Nutch*, созданная для работы с *WARC/ARC*). Хотя крупнейшие поисковики – *Google*, *Yahoo* – утверждают, что способны искать изображения и видео, однако эти опции пока недостаточно отработаны для получения точных результатов на больших массивах.

Каталогизация – это важный фактор, влияющий на поисковые функции сетевого архива. Библиотеки выполняют крупномасштабные работы по архивированию и поэтому стремятся использовать свои традиционные процедуры каталогизации в работе с сетевыми документами, что в последующем

ограничивает возможности поиска.

Традиционная каталогизация не может быть подходящим вариантом для всех видов архивирования; исключение составляет селективное архивирование, поскольку оно обычно по сути своей не слишком масштабно. Только небольшая часть сетевых документов содержит надежные описательные метаданные. Исследования показывают, что каталогизация не будет жизнеспособной опцией для обеспечения доступа к архивированным документам.

В финальном отчёте по проекту «Минерва» (*Minerva Project*) сделан вывод: «Библиотеке следует полагаться на автоматическое индексирование как основной источник и средство обнаружения информации о веб-сайтах и документов в их составе. Для текстовых материалов должен индексироваться полный текст и таким образом обеспечивать пользователям возможность поиска. Библиотеке не следует расширять вложения в МАРК-каталогизацию или в другие ручные методы каталогизации веб-сайтов, за исключением особо интересных для библиотеки и её пользователей сайтов».

Изучение эволюции терминологии. Проект Европейской Комиссии *Living Web Archives* (LiWA) нацелен на улучшение качества сетевых архивов, в частности мультимедийных документов, обнаружения спама за счёт применения инновационных методов фиксирования документов и семантической эволюции.

Одной из задач проекта было стремление проследить эволюцию терминов, что особенно важно при долговременном хранении материалов. По мере изменений в обществе, наблюдаются и изменения в языке – их нужно учитывать при архивировании. Например, при поиске устройств для воспроизведения музыки современный пользователь введёт запрос «*ipod*», хотя ранее использовались бы термины «*discman*» или «*walkman*». Различные изменения происходят и в названии мест, должностей.

Для того чтобы сохранить семантическую доступность сетевых архивов, в проекте *LiWA* предложено создавать кластеры подобных слов. Особенно важен такой подход в архивировании социальных материалов, поскольку они очень быстро эволюционируют, пользователи модифицируют фразы и образуют новые слова.

Библиотека может поддержать разработку интерфейсов веб-сервисов, в которых отслеживались бы изменения в лексике живого Интернета, например на основе *SOAP* (*Simple Object Access Protocol*) – простого протокола доступа к объектам.

Аналитика Сети. Помимо технологических аспектов сетевого архивирования, те организации и учреждения, которые участвуют в создании архивов, должны быть готовыми к ответам на следующие вопросы:

Каким образом пользователи осуществляют поиск нужных им документов (ключевые слова, навигация и т.п.)?

Как отфильтровать плохо проиндексированные или вовсе не индексированные документы?

Какой вид архивных документов пользуется спросом?

Каким образом можно повысить эффективность использования документов, если говорить о поиске по индексированным материалам?

Как можно выявить «образцы поведения» пользователей сетевых архивов?

Подобные вопросы возникают в связи с использованием технологии глубокой обработки, которая осуществляется применительно к материалам «живой» Сети. В данном случае речь идёт об извлечении нового знания из архивированных данных, таких как лог-файлы (файлы регистрации). В этих файлах содержится информация относительно используемых версий программного обеспечения, скорости сетевой передачи, о сетевых серверах, операционных системах, гиперсвязях. Большинство учреждений, ведущих архивирование Сети, не архивируют лог-файлы, хотя их массив позволяет делать глубокий и разносторонний анализ материалов Сети.

Данные глубокого анализа Сети можно сгруппировать следующим образом:

1. Метаданные – их можно дополнительно разделить на две группы – объектные и технические. Объектные метаданные содержат информацию о размере объекта и дате его создания, а технические – несут информацию о типе сетевого сервера и операционной системе.

2. Данные о пользователе – дата запроса к сетевому архиву, объём переданных данных, адрес запроса, IP-адрес пользователя и т.п.

3. Инфраструктурные данные – базовые данные об инфраструктуре системы архивирования (IP-адреса), количество автономных систем и циклов центрального процессора в связи с работой роутера, ширина полосы роутера и т.д.

Всё это улучшает производительность процесса сетевого архивирования и обеспечивает более качественную категоризацию массива и прогноз поведения пользователей.

Проблемы в управлении инфраструктурой

Облачные вычисления и сетевые вычисления. Инфраструктура сетевых архивов должна быть достаточно прочной при работе с колоссальными объёмами данных и способной передавать эти данные пользователю. Основатель Интернет-архива Брюстер Кале (*Brewster Kahle*) так определяет базовые требования к системам архивирования: в системе должно использоваться только обычное оборудование и устройства; система не может полагаться на коммерческие программы; для обслуживания системы не требуется кандидатская степень; система должна быть как можно более простой.

В настоящее время организации, ведущие сетевое архивирование, часто пользуются распределенными системами, в том числе сетевыми (*grid computing systems*). При выполнении сетевых вычислений определяется наличие свободных серверов в Сети, конкретная задача направляется на свободный сервер, причём в очередности вычислений учитывается приоритетность задач; допускается режим параллельных вычислений. Таким же образом конфигурируются системы памяти, образуя «сеть компьютерной памяти».

Сетевые вычисления обладают определёнными недостатками: если нужно использовать 100 процессоров с занятостью 60 минут каждый, то система будет ждать, пока не появится такая возможность. Если хоть один компьютер выйдет из строя, вся система прекратит работу.

Преимущества облачных вычислений. В случае использования технологии облачных вычислений (*cloud computing system*) можно разделить всю задачу на ряд подзадач и одновременно приступить к их решению на нескольких компьютерах, поэтому выход из строя одного сервера не остановит процесс. Очень важно, что в случае отказа каких-либо систем данные можно будет восстановить.

Примером использования облачных технологий может служить Сетевой архив правительственных документов Латинской Америки и Карибского бассейна (*Latin American Government Documents Archives*), в котором участвуют 18 стран. Этот совместный проект с участием библиотек Университета штата Техас накопил уже 6 терабайт информации.

Рекомендации

1. Этические и политические.

Национальная библиотека должна работать над сетевым архивом с максимальной открытостью, объявляя свои планы и делясь опытом; должна разработать этические рамки или создать Совет по этике Сетевого архива.

2. Юридические проблемы, авторское право.

Законодательство Сингапура по авторскому праву нуждается во внесении поправок, которые должны обеспечить: юридическую определённость сбора материалов в Сети, долговременную сохранность сетевых документов и их использование в интересах науки и культуры; механизмы, облегчающие получение разрешений от правообладателей на повторное использование и трансформацию защищённых работ.

3. Технические проблемы.

Сетевой архив должен наращивать доступ к содержанию отдельных сайтов через использование метаданных. Следует использовать такие инструменты, как *Taverna,* или разрабатывать новые – для улучшения технологического процесса.

Архив должен приступить: к глубокой обработке лог-файлов архивированных материалов, чтобы улучшить качество сбора материалов в Сети; к созданию инструментов по визуализации существующих метаданных; к изучению эволюции терминологии – нужно обдумать включение лингвистических слоёв в архитектуру сетевого архива.

В будущем для селективной архивации некоторых событий WAS может приступить к использованию облачных технологий.

Перевод А. И. Земскова