УДК 026.06

Рейнхард Альтенхонер

Библиотеки как посредники в обеспечении хранения и обслуживании электронными научными данными: практические выводы из проекта Немецкого научного фонда

Цель проекта «Электронная сохранность в библиотеках» – разработка кооперативного сервиса для обслуживания и обеспечения сохранности электронных документов. Проект реализуется как единая сетевая модель.

Доклад на заседании «Роль библиотек в обработке, обеспечении сохранности научных данных и обслуживании ими: международные аспекты», организованном Секцией научно-технических библиотек в ходе 78-й Генеральной конференции ИФЛА (9–16 авг. 2012 г., Хельсинки, Финляндия).

Публикуется с разрешения автора и одобрения аппарата ИФЛА.

Ключевые слова: университетские библиотеки, Германия, информационное обслуживание, научные данные, доступ, поиск, сохранность, архивирование, метаданные.

В 2006 г. на Немецкую национальную библиотеку (далее – ННБ) была возложена обязанность комплектовать и архивировать электронные публикации, что зафиксировано в пересмотренной в том же году версии Закона о национальной библиотеке от 1969 г. Закон предусматривает долговременное архивирование, т.е. необходимость сохранять доступность электронных объектов. Это выдвигает ряд требований по отношению к конкретным организациям, и изучать эти требования следует самым серьёзным образом.

На основе нового Закона, который назначил ННБ ответственной за сбор и обеспечение сохранности электронных публикаций в целом, библиотека за сравнительно короткий промежуток времени обязана увеличить свою ёмкость и изменить ряд технологических процессов.

Исходя из этого, ННБ инициировала специальный проект DP4lib (*Digitalpreservationforlibraries*) и предпринимает дальнейшие усилия для того, чтобы передать другим учреждениям свои наработки. Кроме ННБ в этом проекте участвуют: библиотека Университета г. Геттинген и земли Нижняя Саксония (*Niedersächsische Staats und Universitätsbibliothek Göttingen, SUB*), Центр библиотечного обслуживания земли Баден-Вюртемберг (*Bibliotheksservice-Zentrum Baden-Württemberg, BSZ*), Немецкий институт международных исследований в образовании (*GermanInstituteforInternationalEducationalResearch, DIPF*), Головной офис Единой библиотечной сети (*HeadOfficeoftheUnionLibraryNetwork, VZG*), библиотека Университета г. Дрезден и земли Саксония (*SaxonStateandUniversityLibraryDresden, SLUB*), Немецкая национальная библиотека по науке и технике (*GermanNationalLibraryofScienceandTechnology, TIB*), библиотека Университета г. Йена и земли Тюрингия (*Thüringer Universitäts und Landesbibliothek Jena, ThULB*).

Расширение обязанностей не было неожиданостью для ННБ: к 2006 г., когда был принят новый Закон о Немецкой национальной библиотеке, мы уже могли обосновать ряд действий на основе той подготовительной работы, которую выполнили за предыдущие годы.

Начиная с 1990-х гг. ННБ вела предварительные проектные исследования для наработки опыта и создания специальных сервисов по сбору электронных объектов различного типа.

Эти исследования не были специально нацелены на сохранение электронных документов, однако они помогли понять потребности и специфические требования к процессам сбора электронных объектов.

Большинство проектов, выполненных в то время, создавались под определённые типы объектов, например, «Проект по комплектованию газет» или «Специальный проект по приёму продукции электронных журналов, выпускаемых издательством Шпрингер», а также проект создания системы сбора электронных диссертаций и авторефератов.

Все эти проекты позволили приобрести достаточный опыт, в том числе в области метаданных и проработки технологических процессов. Было признано необходимым организовать постоянную идентификацию электронных объектов в цифровой среде, поэтому ННБ инициировала создание сервиса для присвоения постоянных идентификаторов на основе единого имени ресурса (*Uniform resource name, URN*), и сегодня эта система охватывает более 7 млн уникальных объектов.

Однако в ходе дальнейшей работы выявилась серьёзная проблема: несколько индивидуально разработанных технологических процессов не были объединены на организационном и технологическом уровне. Это означало, что они могли успешно работать (в ручном режиме) лишь с небольшим количеством объектов. Обрабатывать большие массивы автоматизированным способом они не могли, и для обслуживания этих систем нужно было привлекать в библиотеку специалистов по информационным технологиям. Кроме того, ещё одна особенность этих инструментов создавала сложности: в то время особое внимание было уделено проблемам доступа к объектам, что привело к появлению различных конкурирующих интерфейсов. Полностью отсутствовал один очень важный аспект – проверка и технический анализ документов с учётом их формата и целостности данных.

По сути, ННБ располагала совершенно независимыми программными решениями, которые охватывали широкий круг задач: сбор различных типов объектов, их проверку, хранение и предоставление пользователю. Когда Закон вступил в силу, стало ясно, что такой подход не отвечает задачам будущей деятельности, требуется его существенная переработка, а точнее – совершенно новый подход.

ННБ активно приступила к выполнению обязанностей по сбору, индексированию и предоставлению доступа к электронным объектам. Но проблема сохранности не находила своего решения. Поэтому в 2004 г. была начата разработка системы долговременного архивирования, и в 2006–2007 гг. создан прототип системы, названный KOPAL (*Cooperative Development of a Long-Term Digital Information Archive*). Разработка была основана на коммерческом продукте, созданном компанией IBM, и стандартном программном обеспечении DIAS (*Digital Information Archive System*).

Ключом и вершиной проекта стала подсистема обработки метаданных, сфокусированная на технических метаданных, в том числе на методах контроля и проверки целостности объекта. Эта подсистема подходила для целого ряда цифровых объектов, которые были готовы для приёма в систему долговременного архивирования. Конечно, объект нужно было проверить на предмет его технической исправности и логической совместимости. Такие операции проводились перед комплектованием в библиотеке открытого доступа (*Open Source Library*), которую мы назвали *koLibRI* (*kopal library for Retrieval and Ingest*).

За исключением библиотечного программного обеспечения на основе Java, которое можно было менять и адаптировать к различным сценариям работы ННБ, комплекс архивирования DIAS представляет собой «чёрный ящик» – в том смысле, что компания IBM несёт полную ответственность за дальнейшее совершенствование программы, контроль за изменениями в программе и работу над ошибками. Кроме того, были предприняты первые шаги по

извлечению данных из архива для практической реализации проектов миграции. После этих разработок сложилась следующая ситуация. С одной стороны, мы располагали несколькими различными технологиями комплектования документов, которые можно было использовать только для определённых классов объектов, независимых друг от друга и достаточно сложных в обслуживании. С другой стороны, имелась совершенно независимая система долговременного хранения, которую нужно было связать с рабочими технологиями комплектования. Сложившуюся ситуацию может проиллюстрировать следующий пример.

При передаче различных рутинных операций выяснилось, что усовершенствованная технология по приёму объектов от производителя или издателя не готова к тому, чтобы проверять техническое качество цифровых объектов. Существующая технология работы с цифровыми объектами реализована как набор независимых процессов в составе системы обеспечения сохранности, и эти процессы начинаются после идентификации и физического приёма объекта. Технологически вполне возможно вставить функции проверки в процедуру сбора объектов. Но столкнувшись с проблемой обработки большого количества данных и необходимостью при этом оптимизировать оборудование для работы с различными типами данных, мы обнаружили, что ННБ придётся начинать новые разработки и пересмотреть действующие технологии. Была принята совершенно новая инфраструктура операций по приёму объектов. Конечно, учитывались и проблемы сохранности.

С самого начала разработка инфраструктуры базировалась на идее кооперации с другими партнёрами, что помогло бы сэкономить деньги и ресурсы. Здесь, в частности, нужно назвать уже упомянутую библиотеку *koLibRI*, разработанную кооперативными усилиями и доступную в Сети, а также сеть передового опыта и обмена технологиями *NESTOR*, в которую входят библиотеки, архивы, музеи и другие организации науки и культуры.

Все участники совместной работы понимают, что многие из задач (например, управление рисками в отношении форматов файлов или выбор конкретных технологий обеспечения сохранности) требуют использования опыта и навыков, накопленных в разных организациях. Поэтому была поставлена задача организовать практическое сотрудничество.

С точки зрения ННБ, следует выделить две главные задачи: 1) повышение гибкости и возможности увеличения производительности процессов сбора цифровых материалов и практическая интеграция процессов обеспечения сохранности цифровых документов в технологии ННБ; 2) расширение практической кооперации для повышения продуктивности и сбережения ресурсов.

Решение этих задач мы начали по следующим направлениям: во-первых, концентрация на разработке технологических процессов, во-вторых, распространение опробованной методики сохранения электронных документов на работу других партнёров, с тем чтобы можно было обмениваться опытом и создать гибкие и адаптированные сервисы.

Разработка технологических процессов. Реализация автоматических процессов основана на трёх базовых требованиях: 1) использование стандартных форматов метаданных для спецификации и проверки электронных ресурсов в каталоге или поисковой системе; 2) определение уровня качества для форматов файла, имеются в виду потребности обеспечения сохранности электронных документов; 3) определение интерфейсов трансфера при получении цифровых объектов и метаданных от изготовителя.

Помимо создания метаданных и налаживания системы контроля качества объектов, изучался процесс передачи объектов и метаданных в ННБ. Сейчас ННБ имеет три интерфейса для поставок: сетевая форма для одиночных объектов и два автоматизированных метода: один – для получения (приёма) объектов, специально направленных в ННБ, другой – для отбора объектов, выставленных в свободном доступе.

Метод приёма использует инструмент доставки (условное название – «горячая папка»), при котором передача осуществляется с применением протокола SFTP или интерфейса WebDAV. Каждый пакет доставки представляет собой единый транспортный контейнер, в котором упакованы как сам объект (он может состоять из многих файлов), так и связанный с ним набор метаданных.

Метод выборки нужных объектов из открытых хранилищ основан на протоколе OAI-PMH в комбинации с трансферными адресами URL, поставляемыми в составе метаданных.

Весь процесс задокументирован как текстовый документ с использованием методики графов BPMN (*Business Process Model and Notation*), помогающих визуализации рабочего процесса и выявлению потребности в технических компонентах.

Дополнительно были предприняты определённые шаги по систематизации сбора сетевых публикаций. Сформирована рабочая группа из сотрудников нескольких отделов. Работа фокусировалась на взаимодействии с учреждениями, а не на сборе отдельных объектов. Таким образом вовлекались агрегаторы и сервис-провайдеры, а затем в группу включили поставщиков программного продукта. Разработка программного обеспечения сама по себе была определена как специальное направление проекта и фокусировалась на производительности и больших объёмах обработки.

Обеспечение сохранности электронных документов для библиотек

Цель проекта DP4lib, финансируемого Немецким научным фондом (2010–2012 гг.), в котором участвуют восемь партнёров, – организационное и техническое расширение проекта KOPAL и превращение его в интегрированный сервис, способный удовлетворять различные потребности партнёров в обеспечении сохранности цифровых данных на единой технологической базе.

По степени активности участники различаются: одни учреждения желают действовать только как пользователи и целиком полагаются на партнёров, создающих готовый продукт; другие хотели бы играть активную роль и отвечать за конкретные задачи. Всё это должно быть сбалансировано и интегрировано в единый коллектив со своей организационной структурой, где учитывается вклад каждого из участников. Построение технической инфраструктуры, разработки на уровне программ и управления цифровыми потоками переданы внешним исполнителям.

Таким образом, проект был сфокусирован на кооперации различных сервис-провайдеров, создании совместных сервисов и организационной модели для долговременного архивирования.

Основная идея состояла в том, чтобы адаптировать и совместно обеспечивать сохранность электронных документов в кооперативной среде. На практике необходимо было чётко прописать и формализовать взаимоотношения между различными группами, с тем чтобы обеспечить работу сервисов, функционально интегрированных в модель обслуживания. В дополнение к технической реализации проекта потребовалось уточнение структуры процессов и экономической модели.

Чтобы создать общий проект, который мог бы реализоваться в самых разных обстоятельствах и интегрироваться в устоявшиеся технологии учреждений сферы культурного наследия, была обозначена цель – использовать открытую концепцию с сервисами модульной структуры. В соответствии с этой целью поставлен ряд задач, среди которых: создание гибкой инфраструктуры системы долговременного обеспечения сохранности, адаптированной к потребностям учреждений культурного наследия и их сервис-провайдерам; техническое

усовершенствование системы KOPAL с учётом интересов всех участников; реализация модели с возможностью повторного использования процессов и подготовка справочника в помощь внедрению технологий сохранности.

Принятая в ННБ технология послужила базой для формирования единого и общего для всех участников технологического модуля, который используется другими партнёрами, но при этом остаётся интегрированным в инфраструктуру информационных технологий ННБ.

При выборе партнёров учитывались их функциональная направленность, заинтересованность и наличие опыта работы в организации сохранности электронных документов. Опыт всех участников следовало собрать и обобщить для достижения конечной цели – создания модели кооперативной структуры и выработки существенных для всех требований. Также следовало выработать конкретные рекомендации, которые бы не только содержали технические и функциональные требования по обеспечению долговременного доступа, но также устанавливали оперативные и организационные нормативы.

С самого начала был предпринят систематический анализ требований от различных учреждений, которые послужили основой выработки внутренних программ и норм долговременной сохранности и доступа к объектам. На ранней стадии проекта выяснилось, что вопросы управления качеством, а также рисками (хотя всё это – изначально технические аспекты) имеют огромное значение.

Основой для анализа стал развёрнутый вопросник, который помог сформулировать основные позиции и самооценку. Каждая организация должна была уточнить, какие именно коллекции она готова взять на долговременное хранение, и что она уже сделала для этого, и т.п.

На следующих этапах требования были обобщены и сокращены с тем, чтобы дать ясное и понятное представление относительно необходимых свойств. Некоторые из функциональных требований содержат положения, выходящие за рамки самого проекта долговременного архивирования, и ориентированы, скорее, на чисто технические аспекты. Нас больше интересовали процессы поддержания сервисов, стабильность управления.

Очень важным аспектом в дополнение к определению технологических задач было введение в систему управления качеством и формулирование требований к документации и отчётности. На основе предварительной работы были уточнены и смоделированы рабочие процессы. Ключевые процессы – такие как сбор документов, обеспечение доступа и сохранности – были идентифицированы; затем проводилась дополнительная сегментация по субпроцессам. Это помогало проводить технические разработки. В то же время были определены и включены в проект различные формы отчётности.

Наши договорённости должны рассматриваться не как жёсткая конструкция, а, скорее, как начальная точка для анализа потребностей и проведения необходимых изменений, что означает постоянную ревизию и оценку всего процесса в целом.

Технологические улучшения во многом были основаны на существующей инфраструктуре, которая либо входила в состав системы KOPAL, либо присутствовала как часть системы сбора данных в ННБ. Существующий SFTP-сервер («горячая папка») был расширен, а техническая обработка направлена на существующий интерфейс OAI-PMH.

Для самого рабочего процесса следует убедиться в обеспечении гарантированного уровня целостности цифровых объектов в ходе процесса сбора документов, а также в том, что такая целостность может поддерживаться постоянно на заданном уровне.

Решающий фактор – постоянное документирование всех этапов посредством создания и

доставки машиночитаемых протоколов передачи, формируемых стандартным образом. Они основаны на билетной системе, которая отслеживает все виды действий, относящиеся к индивидуальному пакету.

Ещё одним важным моментом была реализация системы оценки риска на самых ранних стадиях процесса. Для этого были сформулированы требования по поддержанию целостности файлов и реализации процедуры оценки рисков; развита концепция, или принцип, классификации цифровых объектов. Классификация – многоэтапная процедура, и мы считаем, что это гарантирует долговременную доступность цифровых документов.

Такое заявление основано на результатах технической проверки (мониторинге сервиса). Критерии, среди прочего, формулируются с учётом целостности данных («мы получили тот же самый пакет битов, который был отправлен поставщиком данных?»), идентификации, ограничений на доступ (например, защитный механизм), возможности извлечения технических метаданных, касающихся формата файла, и проверки самого формата. На основе результатов проверки автоматически формируется показатель уровня документа, который в последующем может использоваться как элемент долговременного архивирования.

Уровни приёма строятся иерархическим образом. Высокий уровень означает более высокие ожидания в обеспечении сохранности конкретного объекта. Уровень приёма ранжируется от ноля (при архивации гарантируется только безопасная передача данных) до 4, когда учитываются все критерии (например, «документ формата PDF с использованием согласованной процедуры проверен на целостность и улучшен в аналитическом инструменте как не имеющий ограничений»). Даже генератор метаданных может обрабатывать и поставлять информацию по характеристикам формата, отражённым в технических метаданных. В добавление к проверке формата файла может быть подтверждена также спецификация PDF.

Ещё один блок работы над проектом – определение коммуникационной и организационной структур. Он, в частности, предусматривает подробное описание каждого шага сервиса и технологического процесса; необходимо также определить: глубину документации, периодичность отчётов и их состав, цепочку докладов об ошибках, исходные спецификации для организации деятельности по сохранности. Должны учитываться самые разные сценарии, например, каким образом организовать повторную проверку на дублетность всех объектов коллекций, когда-либо полученных от того или иного поставщика.

Были технологически согласованы некоторые интерфейсы коммуникаций, например веб-сервис доступа (построен на основе программы SOAP/REST). Особое внимание уделялось автоматизации связи между партнёрами (процесс обмена данными с использованием протоколов SFTP и OAI), использованию интерфейса подачи для отчётов (хотя можно было пользоваться интерфейсом приёма).

Чтобы убедиться в качестве управления, сохранять прозрачность и понятность процесса, согласован формат ряда отчётов.

Наиболее сложным оказалось планирование этапов процесса обеспечения сохранности. Работа с объектом строилась таким образом, чтобы собрать всю информацию для сохранности. Нужно было согласовать совместную работу и каталог сервисов, который основан на базовых требованиях, сформулированных в OAIS, и модулей в его составе.

Технологическая цепочка выглядит следующим образом. Сначала должны быть определены соответствующие области работы, подготовлены инструменты, коммуникации и документация. Следующий практический шаг – рабочие контакты с проектом NESTOR. Кроме того, модель процесса может быть получена на основе результатов, представленных

рабочей группой «Сохранность электронных документов» в докладе по систематической идентификации и классификации. Всё это существенно расширяет модели OAIS.

Особое внимание уделялось построению организационной структуры по заранее оговорённым условиям. Интеграция процессов сохранности электронных документов в устоявшуюся структуру и рабочий процесс группы учреждений сама по себе является проблемой. Поэтому нужно, чтобы сотрудничающие группы подготовили специальный акт о взаимоотношениях партнёров и, в особенности, между сервис-провайдерами и потребителями. Это предполагает целый ряд практических договоренностей. Во-первых, наличие процедуры отчётности по выполненным сервисам. До того как согласиться с оплатой сервиса, нужно подготовить надёжные и прозрачные процедуры оценки (в денежных терминах), особенно для того, чтобы удовлетворять конкретные запросы отдельных учреждений.

Требуется также контрактное закрепление отношений между различными партнёрами, в том числе и чёткое описание задач, обязанностей и процедуры принятия решений. Этот порядок предусматривался не только для взаимоотношений между партнёрами, но и внутри одного учреждения. Такое распределение ответственности предполагает наличие чёткой экономической модели, которая постоянно адаптируется к меняющимся условиям и ценовым факторам. Планируется передавать различные сервисы от одного участника другому (например, подготовка метаданных включена в оценку сервисов обслуживания).

Нынешнее состояние проекта DP4lib. Техническая реализация завершена. Совместно с партнёрами успешно проведено испытание. Подготовлена интеграция внешней платформы для хранения документов. Организационные подходы, в том числе экономические аспекты, контракты, отчётность, получили полное одобрение. Пересмотрена технология, введены новый порядок обработки и новая система управления.

Переход к новой системе был тщательно подготовлен, персонал обучен. После практического освоения сервисов в кооперации с несколькими участниками в 2013 г. сервис будет расширен, подключатся новые партнёры, и шаг за шагом будет осуществлён переход к регулярному обслуживанию.

Может показаться, что совместное использование оборудования для долговременного архивирования силами различных учреждений – это простое решение, поскольку все участники заняты одним и тем же делом. Однако на практике оказывается, что различные учреждения могут иметь одинаковые программы, но использовать их по-разному. Кроме того, совместное обслуживание требует от участников ясного понимания своих обязанностей, задач и своей роли в конкретных процессах.

Наш систематический подход состоял в создании модели, учитывающей потребности и нужды организаций глубоко и систематично.

Введение сервисов, обеспечивающих сохранность электронных документов, в практику работы учреждений с уже устоявшимися рабочими технологиями, воздействует на структуру процесса, что ведёт к необходимости взаимной адаптации, приспособления. Если в работе участвуют не только внутренние, но и внешние организационные единицы (а этого зачастую трудно избежать в таких сложных процессах, как обеспечение долговременной сохранности), то важность формирования и глубокого обдумывания сервисов становится решающей.

Создание взаимовыгодной сети учреждений обычно начинается с постановки общих задач – снижения расходов за счёт кооперации, сокращения времени разработок, использования чужого опыта или путём распределения рисков. Затем следует определение отдельных сервисов, соответствующих функций и их реализация. Таким образом, чтобы начать

совместные операции по обеспечению долговременной сохранности, требуется решение в рамках определённой организационной структуры. Потребности, сформулированные теоретически, должны обрести практическое воплощение.

Проект DP4lib– это первый шаг к реализации ориентированной на сервис инфраструктуры для обеспечения сохранности электронных документов. В ходе реализации проекта мы поняли, что намного более важными, чем технический инструментарий, являются разработка технологического процесса, обеспечение качества, надежность, предсказуемые цены, достоверная отчётность и документирование всех процессов.