Формирование тематических научно-образовательных электронных коллекций: интеграция данных и координация комплектования♣

© А.В. Чугунов

Междисциплинарный центр Института филологических исследований
Санкт-Петербургского государственного университета
avc@icape.ru

Аннотация

Статья посвящена рассмотрению организационно-технологических аспектов процесса формирования фонда электронных документов тематической научно-образовательной коллекции. Представлена технологическая схема обработки электронного документа на трех стадиях: а) принятие решения о комплектовании, б) создание электронного документа и определение его статуса в ЭБ, в) размещение документа и технология взаимодействия с другими ЭБ.

1 Региональные ресурсные центры как инструмент координации комплектования электронных коллекций

Известно, что Интернет является децентрализованной структурой без явной географической привязки информационных ресурсов, но сама деятельность происходит, как правило, в конкретных организациях — университетах, академических структурах, библиотеках, органах государственной и региональной власти и т.п. Поэтому при переводе в электронную форму артефактов реального мира — в том числе книг и других публикаций — мы неизбежно вынуждены решать большой комплекс организационно-технических и правовых вопросов.

Идея создания региональных ресурсных центров не является оригинальной. Она довольно часто возникает, когда ставится задача интеграции региональных или ведомственных информационных ресурсов. Обычно такие проекты инициируются в рамках министерств и ведомств, в качестве последнего примера можно привести программу создания таких центров в рамках Минобразования России.

В этой связи следует отметить, что задача создания системы децентрализованного, но скоординированного пополнения русскоязычного электронного документного пространства электронными копиями печатных изданий, носит межведомственный и междисциплинарный характер. Необходимость интеграции и кооперации уже стала достаточно очевидной, и в настоящее время идет процесс институционализации общероссийской ассоциации на базе Некоммерческого партнерства «Электронные библиотеки» (НП ЭЛБИ) [1], которое было создано в феврале 2005 г. (*http://www.elibra.ru*). В число основных направлений деятельности и программы действий НП ЭЛБИ входят проекты по разработке и внедрению реальных механизмов координации и комплектования электронных коллекций, в том числе на основе региональных ресурсных центров. Базовым партнером НП ЭЛБИ на Северо-Западе является Партнерство для развития информационного общества на Северо-Западе России (*http://www.prior.nw.ru*); Северо-Западный ресурсный центр ассоциации создается как межсекторная программа действий в сотрудничестве с Междисциплинарным центром ИФИ СПбГУ, ОАО Линукс-Инк, ОАО Альт-Софт, и другими партнерами.

В результате обсуждения на конференциях, семинарах и круглых столах были сформулированы перспективные направления деятельности Северо-Западного ресурсного центра Российской ассоциации электронных библиотек [2; 3]. В данной публикации мы подробнее остановимся на организационно-технологических аспектах процессса формирования фонда электронных документов тематической научно-образовательной коллекции.

Программа деятельности ресурсного центра предполагает ведение реестра электронных библиотек и коллекций, формируемых на Северо-Западе России и формирование регионального плана оцифровки печатных изданий. Важным аспектом этой деятельности является то, что она предполагает тесную координацию с общероссийским сводным планом оцифровки, который в настоящее время создается в рамках проектов Российской государственной библиотеки и НП ЭЛБИ.

Труды 8ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2006, Суздаль, Россия, 2006.

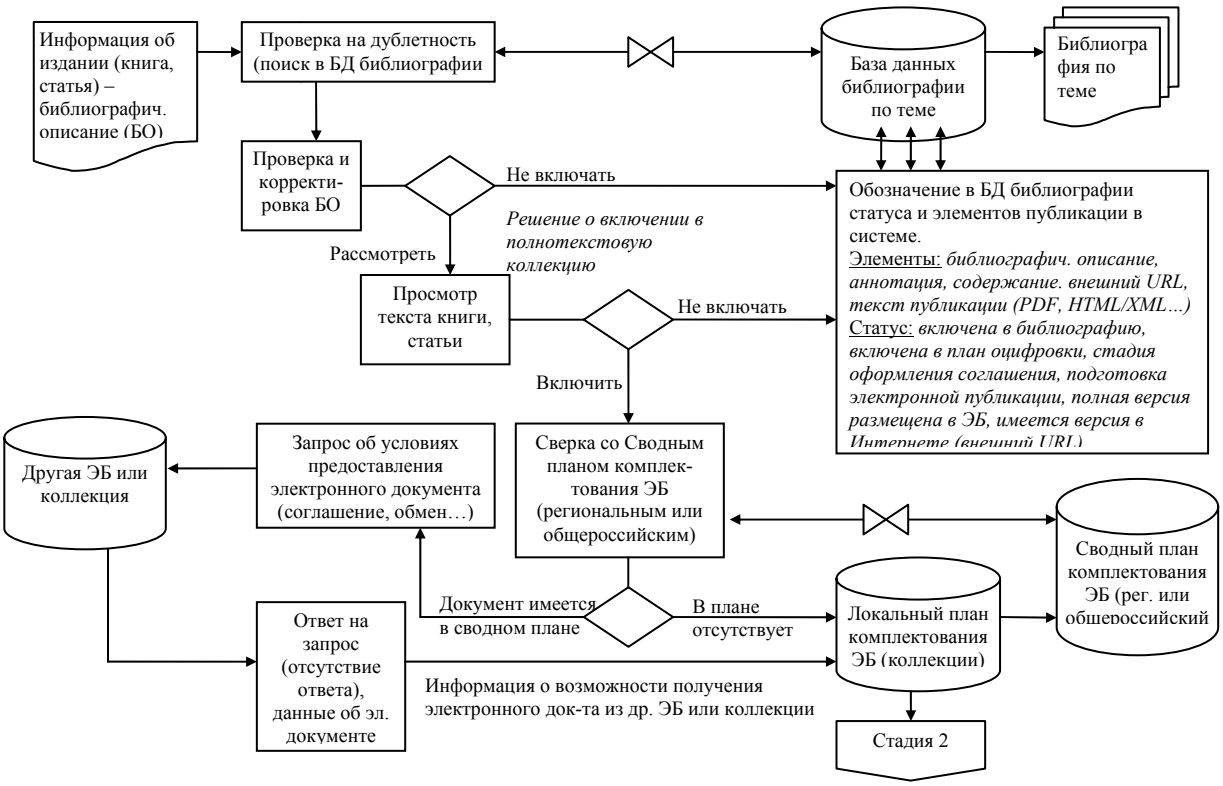

Рис. 1. Технологическая схема обработки электронного документа (стадия 1. План комплектования тематической электронной коллекции)

2 Технология комплектования тематической электронной коллекции

Рассмотрим модель и технологическую схему формирования тематической научно-образовательной электронной коллекции в структуре регионального ресурсного центра. Естественно, сейчас мы можем предложить обобщенную модель и общую логистику информационных потоков и организационных действий. Отработка предлагаемой модели осуществлялась в ходе работ по проектированию, пополнению и развитию двух тематических электронных коллекций: «Электронная библиотека по вопросам развития электронного документного пространства» (*http://library.elibra.ru*) и «Развитие технологий информационного общества» (*http://library.infosoc.ru*).

На наш взгляд, исходным пунктом и началом создания полнотекстовой тематической электронной коллекции должна быть подготовка (или использование уже имеющейся) библиографии по теме. Именно наличие достаточно большого массива библиографических описаний позволит создать (или модифицировать имеющийся) классификатор, описывающий данное предметное поле. Постоянно пополняемая библиографическая БД является основой для формирования плана комплектования проблемно-ориентированной или тематической электронной коллекции. Наличие такого плана позволяет наладить взаимовыгодное сотрудничество с электронными коллекциями смежной тематики, региональными и общероссийскими проектами. Описание детальной логики этого взаимодействия будет формироваться по мере развития регионального ресурсного центра и создания общероссийского сводного плана оцифровки и реестра электронных коллекций.

На рис. 1 представлена первая стадия — принятие решения о включении документов в состав тематической коллекции. На рисунке обозначены связи со сводным (региональным и/или общероссийским) планом комплектования электронных библиотек и коллекций. Следует отметить, что именно возможность получения электронных документов, уже созданных коллегами, для помещения в тематическую проблемно-ориентированную коллекцию и составляет основу мотивации разработчиков коллекций взаимодействовать с ресурсными центрами.

3. Правовой статус и уровень доступа к электронной версии документа

Технологическая схема процесса создания электронного документа и определения его статуса в тематической электронной библиотеке/коллекции (см. рис. 2) описывает вторую стадию ее формирования, которая начинается после того, как описание документа включено в план комплектования и наступила очередь создания его электронной версии.

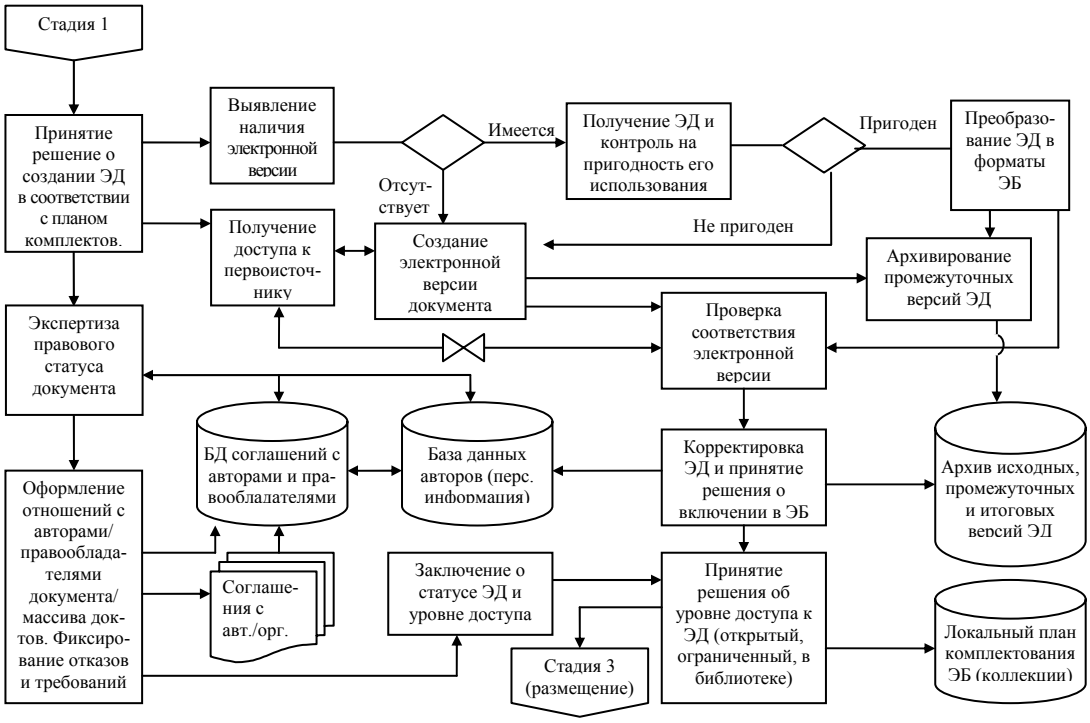

Рис. 2. Технологическая схема обработки электронного документа (стадия 2. Создание электронного документа и определение его статуса в ЭБ)

Важным аспектом этого кооперативного проекта является качество электронных документов. Возможны ситуации, когда есть все составляющие: нужный электронный документ имеется в партнерской коллекции, соглашение о сотрудничестве достигнуто, возможность его передачи в другую тематическую коллекцию существует, документ по запросу предоставляется. Однако использование документа оказывается затруднено или невозможно без существенной модификации. Причины достаточно простые: от несовпадения базовых форматов представления в этих двух коллекциях до выявления ошибок и небрежностей в изготовлении электронной версии документа. В этой связи можно подчеркнуть важность задачи определения набора базовых форматов и конвенциально-принятых сообществом стандартов представления электронных документов. Причем это может быть достаточно широкий список, что позволит учесть интересы большинства электронных библиотек и коллекций.

Представляется, что первым этапом необходимо запустить три задачи: выявление наличия электронной версии (размещена в Интернете, имеется в издательстве, у автора, в организации и т.п.); получение доступа к первоисточнику (для последующей сверки или сканирования); экспертиза правового статуса этого документа (книги, статьи, тезисов доклада и т.п.).

Иногда эти задачи могут быть решены достаточно быстро и эффективно — например, когда мы имеем дело с тезисами докладов на конференциях: на сайте конференции имеется полный текст доклада (и/или в оргкомитете хранится файл печатного сборника); организаторы данной конференции предупреждали авторов докладов о том, что материалы, принятые к публикации, будут представлены в Интернете и могут быть переданы в научно-образовательные электронные коллекции для некоммерческого использования (аналог публичной оферты); имеется соглашение с организаторами конференции и сборник докладов доступен для сверки. Существенным дополнительным плюсом может быть соблюдение организаторами конференций технологии, когда в Интернете тексты докладов размещаются параллельно с выходом печатного издания, являются точной его копией (формируются из оригинал-макета после корректуры и правки) и содержат ссылку на полное библиографическое описание с указанием страниц в бумажной версии. Такой подход незначительно усложняет процесс подготовки к размещению в Интернете, но существенно повышает возможности использования текстов в научной работе и снижает трудоемкость размещения этих текстов в тематических электронных коллекциях. С примером реализации можно познакомиться на сайте Всероссийской конференции «Интернет и современное общество» (сборники трудов конференции за 2004 и 2005 гг. – *http://conf.infosoc.ru*).

Естественно, реализация предлагаемой технологии связана с существенными трудностями, особенно в случаях, когда речь идет о монографиях и учебниках, где, кроме авторов, в качестве правообладателей выступают издательства и/или издающие организации. Особенно тяжело договариваться с издательствами в тех случаях, когда имеется еще не распроданный тираж книги. Готовность авторского коллектива к сотрудничеству и их разрешение

на представление книги в открытом доступе в Интернете не может считаться достаточным основанием, особенно в тех случаях, когда имеется авторский договор, где обозначены позиции с учетом интересов издательства. Имеются и исключения, например, в рамках издательской программы Интернет-университета информационных технологий уже более 20 учебников размещены в Интернете в свободном доступе параллельно с их распространением в торговой сети [4]. Характерно, что руководитель этого проекта А.В. Шкред утверждает, что наличие полной версии книги в Сети не препятствует, а напротив, стимулирует приобретение самих учебников.

Следует отметить, что деятельность ресурсных центров может оказать существенную помощь создателям научно-образовательных электронных коллекций в юридическом оформлении отношений с правообладателями. Важно, чтобы это происходило в рамках проектов, имеющих отдельное финансирование, размер которого позволяет привлекать юридические и консалтинговые компании, способные оперативно реагировать на изменение российского и международного законодательства в сфере авторских прав и информационного права.

Важными аспектами является оформление и распространение типовых договоров по цепочке правоотношений «автор — издатель — держатель ресурса — провайдер» и методическая помощь партнерам в оформлении подобных договоров. Эта деятельность в настоящее время осуществляется при методической помощи Российской ассоциации электронных библиотек (НП ЭЛБИ) [5].

За пределами представленной схемы остаются описание этапов процесса оформления отношений с правообладателями и стадий получения разрешений на размещение текстов в тематической научно-образовательной электронной коллекции. Описание технологии организационно-правового сопровождения создания электронных коллекций может быть темой отдельной публикации. В данном случае можно только констатировать, что вопросы соблюдения авторских прав решаются тем или иным способом. В случае отсутствия таких разрешений электронный документ может использоваться с учетом ограничений, накладываемых действующим законодательством. Вторая стадия завершается принятием решения о размещении электронного документа в коллекции и определением уровня доступа пользователей (открытый или с ограничениями).

4. Размещение документа в ЭБ и взаимодействие с внешними информационными системами

Завершающая стадия размещения документа в электронной коллекции состоит из нескольких этапов, каждый из которых связан с определенными операциями, последовательность которых представлена на блок-схеме (см. рис. 3). Важным вопросом, который ставят для себя все создатели электронных коллекций, является принцип генерации имени файла электронного документа. Чаще всего название файла формируется с использованием фамилии автора текста или фрагмента названия документа, что вызывает серьезные проблемы, связанные с необходимостью создания специальных средств регистрации этих названий и исключения появления дублирующих имен. На наш взгляд оптимальным решением является ведение списка инвентарных номеров электронных документов и присвоение файлу имени повторяющего этот порядковый номер. При этом возникает вопрос — как поступать с документами большого объема, которые следует представлять в виде нескольких файлов (например, книга с разделением на главы или разделы)? Для таких документов нами было принято решение считать часть документа (глава, раздел, параграф, представленные отдельными файлами) отдельной единицей учета и присваивать этому файлу свой инвентарный номер. При этом решение о разделении одного документа (книги, сборника и т.п.) на части принимается с учетом возможности использования каждой части как самостоятельной смысловой единицей. Например, монография имеющая пять глав делится на шесть файлов, где первый включает титульный лист, предисловие, оглавление, введение и т.п., второй файл включает первую главу и далее. Из оглавления ставятся гиперссылки на пять глав (файлов). При этом на каждый файл составляется отдельное библиографическое описание и готовится набор метаинформации (коды рубрикатора, ключевые слова и др.). Такой подход позволяет существенно увеличить количество «точек доступа» к документу.

Важным моментом является информирование профессионального сообщества о новых поступлениях в тематическую электронную коллекцию. Решение этой задачи возможно как традиционными способами (списки рассылки, информационные бюллетени и т.п.), так и чрез взаимодействие с системами, обеспечивающими текущее информациионное обслуживание. Примером такой системы может служить система Соционет (*http://socionet.ru*), предоставляющая возможности онлайнового взаимодействия профессионального сообщества, обеспечивающая ряд информационных сервисов, в том числе так называемого «персонального информационного робота». Этот сервис позволяет пользователю сформировать один или несколько персональных профилей — фильтров отбора описаний документов определенной тематики [6].

Отработка взаимодействия тематической полнотекстовой электронной коллекции с информационной системой поддерживающей функционирование профессионального информационного пространства осуществляется в рамках проекта «ИнфоЛиб: Формирование открытого профессионального сообщества разработчиков научно-образовательных электронных коллекций на базе технологий Соционет» (*http://www.infolib.ru*).

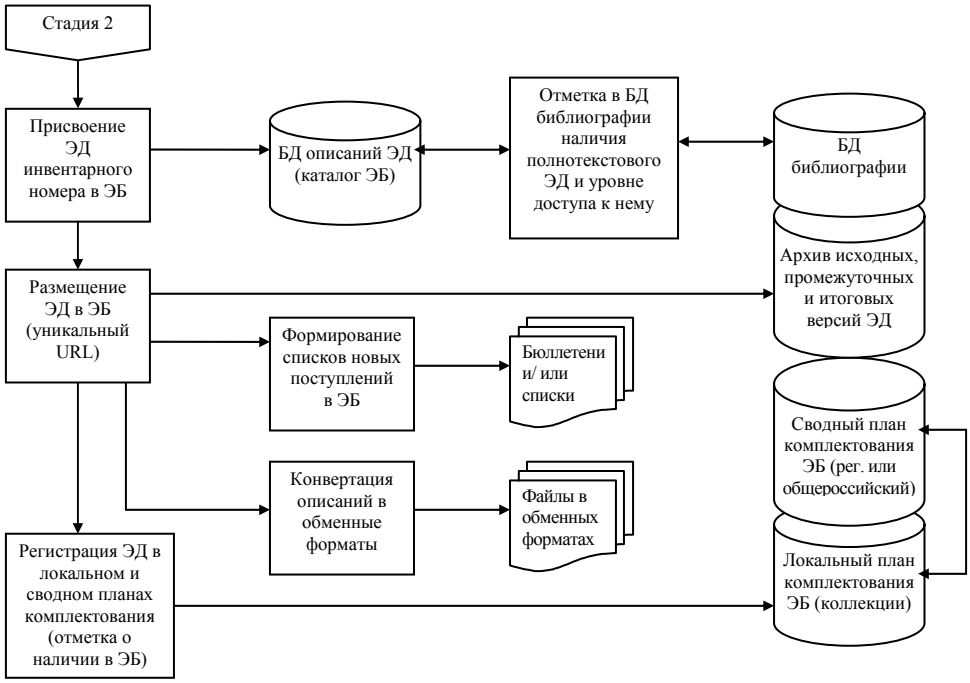

Рис. 3. Технологическая схема обработки электронного документа (стадия 3. Размещение электронного документа в ЭБ)

Для обеспечения текущего взаимодействия и передачи файлов описаний электронных документов из электронной коллекции «Электронное документное пространство» в систему Инфо-Либ выбрана технология синдикации новостей (RSS/XML).

Конвертор, взаимодействующий с БД описаний электронных документов, обеспечивает формирование файлов двух обменных форматов:

– в соответствии со стандартным набором полей RSS (обычно используется для экспорта новостных лент);

– XML-файл с тегами обменного формата системы Соионет/СИНИН, представляющий собой расширение RSS, адаптированное под специфику описания документа более сложной структуры, чем новостные ленты.

Созданная система поддерживает возможность реагирования на запросы к базе данных и отбор документов введенных (или отредактированных) начиная с заданной даты и представление информации в одном из заданных форматов (RSS или XML). В результате выполнения запроса формируется файл с расширением *.rss или *.xml, который обрабатывается специальным скриптом, обеспечивающим загрузку данных в базу системы Инфолиб, реализованной на основе технологических решений Соционет/СИНИН.

Заключительным этапом в технологической цепочке является регистрация электронного документа (с указанием его URL, статуса и названия тематической коллекции) в локальном и сводном планах комплектования. Тем самым создается возможность обеспечить не только текущее планирование формирования электронных версий документов, но и создавать разнообразные информационные сервисы, позволяющие коллективам реализующим проекты, направленные на развитие тематических электронных коллекций, оперативно получать информацию о новых поступлениях в ЭБ смежной тематики. Создание системы региональных и связанного с ними общероссийского плана комплектования ЭБ и соответствующих информационных сервисов позволит постепенно построить эффективный механизм интеграции научно-образовательных электронных коллекций.

Заключение

Предложенную технологическую схему формирования тематической научно-образовательной электронной коллекции можно рассматривать как обобщенную модель и описание информационных потоков, а также соответствующих организационных действий. Вполне естественно, что эта схема не отражает всего разнообразия ситуаций, сопровождающих процесс формирования электронных коллекций.

Коллектив Междисциплинарного центра Института филологических исследований СПбГУ в настоящее время выполняет несколько проектов, направленных на создание научно-образовательных информационных ресурсов гуманитарного профиля, имеющих в своем составе полнотекстовые, графические и мультимедийные коллекции. Поэтому вопросы интеграции тематически близких ресурсов

и проблемы кооперации при создании полнотекстовых и библиографических электронных коллекций становятся весьма актуальными.

Технология и методические решения, представленные в настоящей публикации применяются для пополнения двух тематических коллекций:

– «Электронная библиотека по вопросам развития электронного документного пространства» (*http://library.elibra.ru*);

– «Развитие технологий информационного общества» (*http://library.infosoc.ru*).

В рамках проекта РГНФ «Создание Северо-Западного регионального сервера сопровождения научных электронных коллекций в гуманитарной сфере» осуществляется проработка логистики интеграции коллекций, в частности для создания сводного каталога электронных библиотек и коллекций в области этнографии и смежных дисциплин.

Автор заинтересован в сотрудничестве с коллегами, реализующими проекты создания электронных коллекций в гуманитарной сфере.

Литература

[1] Антопольский А.Б. Формирование электронного документного пространства в России: проблемы взаимодействия и кооперации создателей электронных коллекций / А.Б. Антопольский, Т.В. Майстрович, А.В. Чугунов // Информационные ресурсы России. 2005. № 1. С. 2 – 5.

[2] Чугунов А.В. Создание Северо-Западного ресурсного центра Российской ассоциации электронных библиотек и задачи интеграции научно-образовательных информационных ресурсов / А.В. Чугунов // Труды XII Всероссийской научно-методической конференции Телематика'2005. Т. 1. СПб., 2005. С. 310–312.

[3] Чугунов А.В. О программе создания Северо-Западного ресурсного центра Российской ассоциации электронных библиотек/ А.В. Чугунов // IT-инновации в образовании: Материалы Всероссийской научно-практической конференции (27 – 30 июня 2005 г.) / ПетрГУ. Петрозаводск, 2005. С. 251–253.

[4] Шкред А.В. Модель дистанционного обучения / А.В. Шкред // Технологии информационного общества — Интернет и современное общество: труды VII Всероссийской объединенной конференции. СПб., 10 – 12 ноября 2004 г. СПб.: Изд-во Филологического ф-та СПбГУ, 2004. С. 111–112.

[5] Правовые рекомендации для создателей и владельцев электронных библиотек / Российская ассоциация электронных библиотек (Некоммерческое партнерство «Электронные библиотеки»); под ред. В.Н. Монахова. М., 2006.

[6] Паринов С.И. Интернет-технологии 2-го поколения: осознанные необходимости / С.И. Паринов // Научный сервис в сети Интернет: Труды Всероссийской научной конференции. Новороссийск, 22 – 27 сентября 2003 г. М.: Изд-во МГУ, 2003. С. 138 – 139.

Development of Subject Academic Digital Collections: Data Integration and Coordination of Acquisition

Andrew V. Chugunov

The paper discusses the organizational and technological facets of the electronic documents acquisition process for the subject academic digital collection. The technological charts of processing of an electronic document are presented. The charts cover three stages of the processing: (a) making the acquisition decision, (b) obtaining the electronic document and determining its' status, and (c) placing the document into the collection and technology of the interoperation of the document with other digital libraries.

Maintenance of two digital collections:

– "Digital Library on Digital Document Space Development" (*http://library.elibra.ru*), and

– "Progress of Information Society Technologies" (*http://library.infosoc.ru*)

employs the technology and the methodology presented in this paper.

♣ Работа выполнена при поддержке Российского гуманитарного научного фонда (проекты 04-03-12026в; 05-03-12319в) и Фонда Форда (проект Инфо-Либ).