Опыт создания и поддержки полнотекстовых баз данных неопубликованных документов

© Авдеева Н.В.
Российская государственная библиотека
avdeeva@rsl.ru

Аннотация

Статья посвящена вопросам создания и поддержки российских полнотекстовых баз данных неопубликованных документов, а также предоставления доступа к ним. Определены основные типы неопубликованных документов. Рассмотрены правовые и технологические аспекты создания и развития полнотекстовых баз данных по таким типам неопубликованных документов, как диссертации, депонированные рукописи и научно-исследовательские, опытно-конструкторские и технологические работы (НИОКР).

Важнейшим источником научной информации является документ (от латинского слова documentum – свидетельство) – информация, зафиксированная специальным образом на материальном носителе, снабженная реквизитами, позволяющими идентифицировать документ в процессе его обработки, поиска, использования и хранения. В широком смысле документ служит средством закрепления и передачи информации, знаний, духовных и материальных достижений человеческого общества. Документ является результатом и предметом духовной и материальной культуры.

По социальному статусу документы подразделяются на опубликованные, неопубликованные и непубликуемые документы.

Опубликованные документы – это документы, прошедшие редакционно-издательскую обработку: книги, брошюры, монографии, сборники, тезисы докладов, периодические и продолжающиеся издания, патенты и авторские свидетельства, препринты, стандарты, нормативно-технические документы, прейскуранты, каталоги, авторефераты диссертаций, рекламные издания. Они предназначены для широкого распространения и тиражируются типографским или каким-либо иным способом.

Неопубликованные документы – это документы, не прошедшие редакционно-издательскую обработку и существующие на правах рукописи: отчеты о научно-исследовательских работах, диссертации, описания алгоритмов и программ, проекты, сметы, не рассчитанные на широкое распространение. К неопубликованным документам относятся депонированные рукописи.

Непубликуемые документы – рукописные материалы сугубо личного характера (письма, дневники и др.), не предназначенные для публикации, которые со временем могут быть опубликованы [2,9].

В связи с научной потребностью в различных российских организациях стали формироваться полнотекстовые базы данных неопубликованных документов. Но до сих пор решены не все организационные и правовые задачи создания, обработки и использования полнотекстовых баз данных неопубликованных документов, что привело к необходимости обобщить накопленный опыт, упорядочить использование основных терминов документоведения в данной сфере деятельности, а также обозначить принципиальные правовые и технологические аспекты подобных проектов. Ниже приведен обзор крупнейших российских полнотекстовых баз данных неопубликованных документов.

Диссертации

Диссертация – (от лат. dissertatio – рассуждение, исследование) – квалификационная работа на присуждение учёной степени. В России различают диссертации на соискание учёной степени кандидата наук и доктора наук. Требования к содержанию диссертации различаются в зависимости от учёной степени, на которую претендует соискатель, и от научного направления. Общими требованиями являются оригинальность, научная новизна и практическая значимость работы. [12].

Российская государственная библиотека (РГБ) располагает уникальным фондом подлинников кандидатских и докторских диссертаций, защищенных в стране по всем специальностям, кроме медицины и фармации (национальным хранилищем диссертаций по этим направлениям является Центральная научная медицинская библиотека Первого МГМУ

им. И.М. Сеченова (ЦНМБ) Министерства здравоохранения и социального развития Российской Федерации). В соответствии с приказом Всесоюзного комитета по делам Высшей школы СНК СССР все авторы диссертаций должны были сдавать обязательную копию своих диссертации и автореферата в фонд РГБ. В настоящее время фонд диссертационных работ находится в филиале РГБ в отделе диссертаций г. Химки и составляет более миллиона томов.

Для решения основных проблем: сохранения такого огромного фонда, а главное – обеспечения доступа к нему одновременно большого количества читателей, – с 2001 года РГБ было принято решение о создании Электронной библиотеки диссертаций Российской государственной библиотеки (ЭБД РГБ) на основе современных информационных технологий. В 2003 году был оцифрован стартовый пакет диссертаций по наиболее востребованным специальностям: экономические, юридические, педагогические, психологические и философские науки (всего около 28 000 полных текстов). Начиная с 2004 года, состав ЭБД РГБ пополнялся объемом диссертаций по всем специальностям (кроме медицины и фармации), что составляет около 30 000 – включая 20 000 кандидатских и 10 000 докторских – диссертаций в год. В рамках проекта ретроконверсии в 2006 году были оцифрованы все диссертации за 1985 год. А с 2007 года в ЭБД РГБ поступают диссертации по всем дисциплинам, включая работы по медицине и фармации. Электронная библиотека диссертаций Российской государственной библиотеки (http://diss.rsl.ru) содержит более 725 000 полных текстов диссертаций, защищенных в Российской Федерации и на постсоветском пространстве, по всем специальностям Высшей аттестационной комиссии Министерства образования и науки Российской Федерации (ВАК), а также авторефераты к ним. [3,11].

Каталог ЭБД РГБ находится в свободном доступе для всех пользователей сети Интернет. Полные тексты диссертаций и авторефератов представлены в формате PDF (аббревиатура от Portable Document Format) – формат электронных документов, разработанный компанией Adobe Systems. Кроссплатформенность формата создает удобные условия для организации электронного документооборота. Документ в формате PDF может содержать шрифты, графику, мультимедийные элементы, что гарантирует правильное отображение независимо от операционной системы, программного обеспечения и пользовательских настроек конкретного компьютера.

Для организации доступа к ЭБД РГБ в библиотеках организаций открываются Виртуальные читальные залы РГБ (ВЧЗ РГБ), на территории которых доступ к текстам работ возможен с использованием специализированного программного обеспечения, созданного во исполнение Части четвертой Гражданского кодекса Российской Федерации, вступившей в силу с 1 января 2008 года, где указано: «В случае, когда библиотека предоставляет экземпляры произведений, правомерно введенные в гражданский оборот, во временное безвозмездное пользование, такое пользование допускается без согласия автора или иного правообладателя и без выплаты вознаграждения. При этом выраженные в цифровой форме экземпляры произведений, предоставляемые библиотеками во временное безвозмездное пользование, в том числе в порядке взаимного использования библиотечных ресурсов, могут предоставляться только в помещениях библиотек при условии исключения возможности создать копии этих произведений в цифровой форме» [1]. Высшие учебные заведения и другие организации, имеющие в своем составе библиотеку, могут заключить договор с РГБ на создание ВЧЗ РГБ, по условиям которого каждый читатель ВЧЗ РГБ после регистрации может получить свободный (бесплатный) доступ ко всем полным текстам диссертаций и авторефератов через защищенную программу просмотра, поддерживающую возможность полнотекстового поиска внутри каждой работы. На сегодняшний день создано более 500 Виртуальных читальных залов РГБ почти во всех регионах России и в 10 странах СНГ (Азербайджан, Армения, Беларусь Грузия, Казахстан, Кыргызстан, Молдова, Таджикистан, Украина, Узбекистан), а также за рубежом в Республике Иран [5,15].

Одной из составляющих Электронной библиотеки диссертаций РГБ является Открытая электронная библиотека диссертаций (ОЭБД), которая, по сути, стала предшественником ЭБД РГБ. Разработка проекта ОЭБД велась в несколько этапов. На первом этапе (2002-2003 год) был проведен социологический опрос среди читателей Российской государственной библиотеки на предмет того, готовы ли авторы диссертаций к передаче своей работы для размещения на сайте РГБ в свободном доступе. Как показала статистика, практически все опрошенные читатели дали свое согласие и признали необходимость деятельности в данном направлении. На втором этапе велась работа по гранту «Электронная библиотека диссертаций в теледоступе» (при поддержке Российского фонда фундаментальных исследований (РФФИ), проект № 01-07-90310). Третий этап осуществлялся с 2004 г. по гранту «Интегрированная библиотека электронных диссертаций» при поддержке Российского фонда фундаментальных исследований

(РФФИ, проект № 04-07-90154). Работы на этом этапе включали создание отдельного сайта и каталога, а также использование новой технологии – расширенного языка разметки XML (eXtensible Markup Language) [4,10].

Все диссертации и авторефераты были представлены на сайте Открытой Российской Электронной Библиотеки http://orel.rsl.ru в свободном доступе. Однако в марте 2009 года этот сайт был расформирован. В результате проведенных работ руководством РГБ было принято решение о размещении ОЭБД на сайте Электронной библиотеки диссертаций Российской государственной библиотеки http://diss.rsl.ru и о ведении единого каталога. Это можно считать четвертым этапом, продолжающимся и по настоящее время. В состав ОЭБД входят полные тексты диссертаций и авторефератов, предоставленные авторами для размещения в свободном доступе на сайте Российской государственной библиотеки в формате PDF. Каждый ученый, защитивший диссертацию в России (до 1991 г. – СССР), может передать свою работу для размещения на сайте РГБ, заключив лицензионный договор. Работу можно передать по электронной почте или на электронных носителях (CD, DVD, флешкарте и т.д.). Если диссертация или автореферат уже имеются в наличии в каталоге ЭБД РГБ, то автор может не передавать полный текст, а обозначить в договоре перевод его работы в открытый доступ. Поиск диссертаций и авторефератов осуществляется в общем каталоге ЭБД РГБ по всем доступным поисковым признакам простого и расширенного поиска: по ключевым словам и словосочетаниям, по автору, по специальности ВАК и т.д. Каждая диссертация и автореферат имеют признак доступности: для всех диссертаций в открытом доступе (ОЭБД) ссылка на электронный ресурс зеленого цвета. На данный момент ОЭБД содержит более 3 000 полных текстов диссертаций и авторефератов, размещенных в открытом доступе. Подробную информацию о проекте можно посмотреть на сайте ЭБД РГБ по ссылке http://diss.rsl.ru/?menu=about/31/&lang=ru [3,4,11,15].

Открытие полных текстов документов из ЭБД РГБ возможно при использовании одного из видов программного обеспечения – а именно: web-интерфейса для on-line просмотра, программы Acrobat Reader, систем DefView или DVS. Web-интерфейс и Acrobat Reader применяются только для произведений, находящихся в свободном доступе, а системы DefView и DVS позволяют открывать любые документы из ЭБД РГБ, независимо от наложенных ограничений на доступ к полному тексту произведения.

Система защищенного просмотра документов DefView (Defence Viewer – защищенная программа просмотра) является лицензионным программным продуктом, который устанавливается на каждое рабочее место, где осуществляется доступ к полным текстам произведений ЭБД РГБ. Система DefView используется для доступа к текстам ЭБД РГБ в зале Отдела диссертаций (г. Химки) и во всех виртуальных читальных залах РГБ.

Система DVS (Documents View System – система просмотра документов) является web-приложением и не требует установки. Доступ к полным текстам произведений ЭБД РГБ с использованием системы DVS осуществляется через сайт https://dvs.rsl.ru. Для входа в систему каждый пользователь вводит данные своей учетной записи [5].

Сегодня Электронная библиотека диссертаций РГБ дает уникальный шанс для тысяч ученых по-новому реализовать возможности свои и коллектива, снизить стоимость научных исследований, сформировать свои научные взгляды с учетом знаний, наработанных десятилетиями.

Депонированные рукописи

В России депонированием рукописей занимается Институт научной информации по общественным наукам Российской академии наук (ИНИОН РАН), который был создан в 1969 году на основе Фундаментальной библиотеки общественных наук АН СССР. Он приобрел широкую известность благодаря системе научно-информационных изданий (библиографических, реферативных и аналитических), научным исследованиям в различных областях социального и гуманитарного знания, Фундаментальной библиотеке, насчитывающей более 14 млн. единиц хранения, Автоматизированной информационной системе по общественным наукам.

Общие положения депонирования научных работ:

1. Депонирование (передача на хранение) – особая система публикации научных работ (отдельных статей, обзоров, монографий, сборников научных трудов, материалов научных конференций, симпозиумов, съездов, семинаров) узкоспециального профиля, разрешенных в установленном порядке к открытому опубликованию, которые нецелесообразно издавать полиграфическим способом печати, а также работ широкого профиля, срочная информация о которых необходима для утверждения их приоритета.

2. Депонирование предусматривает прием, учет, регистрацию, хранение научных работ и обязательное размещение информации о них в специальных информационных изданиях.

3. Депонирование научных работ осуществляется при наличии согласия автора(ов) и решения ученого, научно-технического советов, а также редакционно-

издательских советов издательств и редакционных коллегий научных журналов и сборников.

4. Авторы депонированных работ сохраняют права согласно законодательству о защите авторского права, но не могут претендовать на выплату гонорара.

5. Депонированные научные работы приравниваются к опубликованным печатным изданиям.

6. По результатам депонирования по запросу автора в его адрес направляется справка о депонировании научной работы.

7. ИНИОН РАН депонирует научные работы по социальным и гуманитарным наукам. Информация о депонированных в ИНИОН РАН научных работах публикуется в библиографическом указателе «Депонированные научные работы» (база данных «Депонированные рукописи» расположена по ссылке http://83.149.253.12/scripts/Rweb.exe?DBNAME=dep&DCNFN=7221&SYSLANG=RU).

Научные работы представляются на депонирование в двух экземплярах на русском языке в печатном виде. На сайте http://inion.isras.ru/index.php?page_id=180 ИНИОН РАН подробно описаны все условия предоставления научных работ; данная услуга для авторов является платной, но научные работы, направленные Учеными советами учреждений РАН, депонируются бесплатно [13].

Фонд депонированных рукописей ИНИОН РАН составляет на сегодняшний день более 60 000 единиц. Ранее ежегодно поступало до 3 000 рукописей на депонирование в год. К сожалению, их количество сейчас значительно сократилось (до 250-300 в год). Продолжается работа по наполнению библиографической базы данных депонированных рукописей, в которой сейчас содержатся описания 15 000 рукописей, поступивших в фонд, начиная с 1994 года. С 2002 года авторам депонированных работ было предложено присылать дополнительно к печатным версиям рукописей их электронные копии. Если рукопись поступала в электронном виде, то она включалась в полнотекстовую базу данных неопубликованных рукописей. На сегодняшний день полнотекстовый архив депонированных рукописей содержит около 5 000 документов, но в связи с вступлением в силу Части четвертой Гражданского кодекса Российской Федерации доступ к этому архиву закрыт с 1 января 2008 года. При этом библиографическая база данных депонированных рукописей по-прежнему находится в открытом доступе на сайте ИНИОН РАН.

Все представление баз данных ИНИОН РАН (на CD и в Интернете) было реализовано на WebIRBIS™. Поисковая система WebIRBIS™ предназначена для многоцелевой обработки больших, в том числе полнотекстовых баз данных, содержащих разнородные документы неограниченной длины с нерегулярной структурой. Система имеет развитые средства поиска, сортировки и вывода информации, обеспечивая гибкость и эффективность технологий информационного поиска. Более подробная информация о системе представлена на сайте ИНИОН РАН http://www.inion.ru/search-help-rus2.html #O_poiskovoj_systeme [1,13].

Научно-исследовательские, опытно-конструкторские и технологические работы (НИОКР)

Научно-исследовательские, опытно-конструкторские и технологические работы (НИОКР) – совокупность работ, направленных на получение новых знаний и их практическое применение при создании нового изделия или технологии.

НИОКР (в английском языке используется термин «Research & Development» (R&D)) включает в себя:

- Научно-исследовательские работы (НИР) – работы поискового, теоретического и экспериментального характера, выполняемые с целью определения технической возможности создания новой техники в определенные сроки. НИР подразделяются на фундаментальные (получение новых знаний) и прикладные (применение новых знаний для решения конкретных задач) исследования.
- Опытно-конструкторские работы (ОКР) и Технологические работы (ТР) – комплекс работ по разработке конструкторской и технологической документации на опытный образец изделия, изготовлению и испытаниям опытного образца изделия, выполняемых по техническому заданию [12].

В соответствии с поручением Президента Российской Федерации от 4 января 2010 года № Пр-22, пункт 1 «Ж», Министерство образования и науки Российской Федерации ведет работы по формированию Единой федеральной базы данных, включающей результаты научно-исследовательских, опытно-конструкторских и технологических работ гражданского назначения, выполняемых за счет средств федерального бюджета, и проектов внедрения новых информационных технологий, выполняемых с использованием государственной поддержки (ЕФБД НИОКР).

Такая база данных собирается и ведется в Центре информационных технологий и систем органов исполнительной власти (ЦИТиС), который как федеральный информационный центр осуществляет формирование и поддержку национального библиотечно-информационного фонда Российской Федерации в части открытых неопубликованных источников

Таблица 1 *Объем фонда ЦИТиС*

Статистика федерального фонда (1982-2012 гг.)	**Всего**	**2009**	**2010**	**2011**	**2012**
Информационные карты диссертаций	671 835	28 260	24 700	24 800	4 000
Информационные карты НИР и ОКР	1 307 811	12 590	15 300	19 600	3 500
Регистрационные карты НИР и ОКР	1 156 862	23 283	21 100	31 000	1 800
Информационные карты алгоритмов и программ (с 1996 г.)	17 079	1 242	2 020	1 101	260
Объекты учета РНТД (с 2007 г.)	10 656	1 650	1 289	3 844	1 034

научной и технической информации – отчётов о НИОКР, кандидатских и докторских диссертаций, описаний алгоритмов и программ. (Постановление Правительства Российской Федерации от 31 марта 2009 г. № 279. Ранее эти функции выполнял Всероссийский научно-технический информационный центр (ВНТИЦ)).

В настоящее время фонд ЦИТиС насчитывает более 7 млн. документов. Ежегодные поступления в ЦИТиС составляют около 100 000 различных документов, отражающих контент научно-технической информации. Обработка такого количества документов требует значительных финансовых и временных затрат.

Поставщиками информации в ЕФБД НИОКР являются организации науки и высшей школы, промышленные предприятия – исполнители НИОКР, диссертационные советы и авторы диссертаций, а также бюджетополучатели – министерства и ведомства, выступающие государственными заказчиками НИОКР.

Система электронного документооборота научно-технической информации в федеральном информационном центре имеет ряд принципиальных особенностей. Прежде всего, это система, рассчитанная на прием, обработку, хранение и распространение больших объемом информации – несколько сотен тысяч документов в год. При этом объем документов колеблется от 3 Кб до 250 Мб. Информация, представленная в документах, не структурирована либо слабо структурирована, документы относятся к различным областям знаний, т.е. фонд политематичен.

Традиционно документы, представляемые во ВНТИЦ, поступали только на бумаге и, пройдя определенные стадии обработки, трансформировались в электронный вид. По мере развития компьютерных технологий и внедрения их в технологический процесс формирования федерального фонда по непубликуемым источникам информации, различные операции по обработке, хранению и распространению информации автоматизировались и модернизировались, создавая основу для системы электронного документооборота.

Было создано интегральное электронное автоматизированное хранилище ВНТИЦ, включающее банк данных государственных контрактов на НИОКР, политематические ретроспективные реферативно-библиографические базы данных по государственной регистрации и учёту НИОКР и диссертаций, а также хранилище полнотекстовых отчётов и диссертаций объем которого представлен в таблице 1, с реализацией организации онлайнового доступа пользователей через сеть Интернет к ресурсам электронного хранилища.

В настоящее время в ЦИТиС разработана система электронного документооборота научно-технической информации, включающая как технологию и средства приема, так и обработку, хранение и распространение реферативной информации.

Прием документов в электронном виде (регистрационных, информационных карт НИОКР, информационных карт диссертаций) реализован с использованием технологии ASP.NET. Доступ к системе осуществляется через сайт ЦИТиС www.citis.ru (ранее был сайт www.vntic.org.ru). Далее принятые документы поступают в технологическую базу, обрабатываются и загружаются в электронное автоматизированное хранилище ЦИТиС. Первоисточники – диссертации, тексты НИР и ОКР – сканируются и также поступают в электронное хранилище [6,14].

Полные тексты документов и библиографические записи к ним можно посмотреть на безвозмездной основе только в читальном зале ЦИТиС (без возможности создания электронной копии) и заказать на печать фрагменты, не превышающие 20% от объема текста. Удаленный доступ к базам данных предоставляется на платной основе на условиях подписки, подробная информация представлена (http://www.rntd.citis.ru/rntd/online.php) на сайте ЦИТиС.

Литература

[1] Гражданский кодекс Российской Федерации (ЧАСТЬ ЧЕТВЕРТАЯ)

[2] Федеральный закон об обязательном экземпляре от 26.11.1994 (Глава I. ОБЩИЕ ПОЛОЖЕНИЯ Статья 1, Основные понятия)

[3] Авдеева Н.В. Электронная библиотека диссертаций Российской государственной библиотеки: история создания и перспективы развития // Информационные ресурсы России. – 2009. - №5 – С. 17-21

[4] Авдеева Н.В., Лавренова О.А. Интегрированная библиотека электронных диссертаций // Информационные технологии, компьютерные

системы и издательская продукция для библиотек: Доклады и тезисы докладов. – М.: ГПНТБ России, 2004. – С. 110-117 ("LIBCOM-2004")

[5] Авдеева Н.В., Чемоданова О.В. Разработка и поддержка программного обеспечения для Электронной библиотеки РГБ // Материалы Восемнадцатой Международной Конференции "Крым 2011": "Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса" – ГПНТБ России, Ассоциация «ЭБНИТ», 2011. http://www.gpntb.ru/win/inter-events/crimea2011/disk/139.pdf

[6] Голосов Ю.И., Брагина Г.А., Пржиялковская М.Н. Электронные документы научно-технической информации в системе ВНТИЦ // Электронные библиотеки; перспективные методы и технологии, электронные коллекции: Труды десятой всероссийской научной конференции RCDL'2008, Дубна, Россия, 2008. http://rcdl.ru/doc/2008/343_344_paper41.pdf

[7] Гончаров М.В. Современное состояние и перспективы развития библиотечных Интернет/Интранет технологий: диссертация на соискание ученой степени кандидата технических наук: 05.25.05. – М., 2002. – 142 с.

[8] Земсков А.И., Шрайберг Я.Л. Электронные библиотеки: Учебник для студентов вузов культуры и искусств и др. высших учеб. заведений/ А.И. Земсков, Л.Я. Шрайберг. – М.: Либерия, 2003. – 352 с.

[9] Золотарева В.И. Основы информационной культуры [Электронный ресурс]: учебно-методическое пособие/ В.И. Золотарева [и др.]. – М.: МИФИ, 2005. http://library.mephi.ru/icb2/book.html

[10] Лавренова О.А. Новый взгляд на проект электронной библиотеки диссертаций // Электронные библиотеки; перспективные методы и технологии, электронные коллекции: Труды седьмой всероссийской научной конференции RCDL'2005, Ярославль, Россия, 2005. http://rcdl.ru/doc/2005/sek4_2_paper.pdf

[11] Avdeeva N. INNOVATIVE SERVICES FOR LIBRARIES THROUGH THE VIRTUAL READING ROOMS OF THE DIGITAL DISSERTATION LIBRARY, RUSSIAN STATE LIBRARY // IFLA Journal – 2010 – Vol. 36, Issue no. 2, p. 138-144

[12] Сайт Википедии http://ru.wikipedia.org/

[13] Сайт Института научной информации по общественным наукам (ИНИОН РАН) www.inion.ru

[14] Сайт Центра информационных технологий и систем органов исполнительной власти (ЦИТиС) www.citis.ru

[15] Сайт Электронной библиотеки диссертаций Российской государственной библиотеки (ЭБД РГБ) http://diss.rsl.ru

Experience of development and support of full-text databases of unpublished documents

Nina Avdeeva

The article is devoted to the issues of development and support of the Russian full-text databases of unpublished documents, and also to the issues of providing access to them. It distinguishes main types of unpublished documents. It presents legal and technological aspects of development of full-text databases on such document types as dissertations, deposited manuscripts and R&D publications.