Аналитическая обработка документов для обеспечения научных исследований и разработок

Analytical Processing of Documents for Research and Development Purposes

Аналітична обробка документів для забезпечення наукових досліджень і розробок

Воройский Ф. С.
Государственная публичная научно-техническая библиотека России, Москва, Россия

F. S. Voroisky
Russian National Public Library for Science and Technology, Moscow, Russia

Воройський Ф. С.
Державна публічна науково-технічна бібліотека Росії, Москва, Росія

Рассматриваются традиционные и новые подходы к аналитической обработке документов в научно-технических библиотеках, ориентированные на информационное обеспечение и справочно-информационное обслуживание ученых, разработчиков объектов науки и техники, преподавателей вузов и других категорий пользователей. На основе анализа информационных потребностей пользователей особое внимание уделяется координатным языкам индексирования и наличию аннотаций и рефератов в библиографических записях. Одновременно ставится вопрос о необходимости использовании систем метаданных для описания полнотекстовых электронных ресурсов.

The author considers traditional and new approaches to analytical processing of documents in sci-tech libraries oriented on dataware and reference-information services of scientists, designers in science and technology, teachers of high schools and other categories of users. Based on the analysis of user information demands the author gives special attention to the coordinate indexing languages and to the presence of abstracts and reviews in bibliographic records. At the same time, the author raises an issue of using metadata systems for describing full-text electronic resources.

Розглядаються традиційні й нові підходи до аналітичної обробки документів в науково-технічних бібліотеках, що орієнтуються на інформаційне забезпечення і довідково-інформаційне обслуговування вчених, розробників об'єктів науки та техніки, викладачів вузів та інших категорій користувачів. На основі аналізу інформаційних потреб користувачів особлива увага приділяється координатним мовам індексування й наявності анотацій та рефератів в бібліографічних записах. Одночасно ставиться питання про необхідність використання систем метаданих для опису повнотекстових електронних ресурсів.

Не думаем, что следует доказывать тот факт, что библиографическая и аналитическая обработка документов должны быть ориентированы на то, что их результаты должны обеспечить удовлетворение информационных потребностей пользователей путем получения ими релевантных (а лучше — пертинентных) документов и данных, необходимых для выполнения разнородных исследовательских и технических работ или разработок. Очевидно также, что удовлетворение этих потребностей происходит в результате индивидуального поиска в локальных и/или удаленных электронных каталогах Интернета, а также через услуги справочно-информационного обслуживания и обеспечения (СИО и ИО), оказываемые информационными службами опять же через поиск, осуществляемый в этих источниках.

Характерной особенностью СИО является его преимущественная ориентация на выявленные устойчивые или длительно существующие информационные потребности определенных групп пользователей информации. Реализация СИО предполагает выполнение библиотеками и информационными органами достаточно стандартизованных в рамках организаций видов работ по комплектованию справочно-информационных фондов, их каталогизации, созданию и ведению баз данных, поиску и распространению информации по заявленным в форме «запросов» или «подписки» на обслуживание потребностям пользователей и т. п. В отличие от справочно-библиографического обслуживания (СБО), ориентированного на предоставление пользователям (в том числе читателям) сведений библиографического характера, СИО распространяется на подготовку и выдачу заинтересованным лицам и организациям данных любого вида. В указанном контексте СБО можно рассматривать как одну из разновидностей СИО.

По своим основным признакам СИО может быть отнесено к категории сравнительно недорогих «массовых» или «стандартных» видов услуг. Оно не предусматривает возможности удовлетворения информации потребностей «слишком привередливых» или не вписывающихся в общий ряд «сложных» клиентов, нуждающихся в индивидуальной подготовке документов и данных, а также в специальном порядке и сроках их предоставления.

Необходимость устранения указанного недостатка привела к появлению другого режима и связанного с ним понятия — «информационное обеспечение» (ИО), представляющее собой совокупность процессов по подготовке и предоставлению специально подготовленной научно-технической информации (НТИ) для решения управленческих и научно-технических задач в соответствии с этапами их решения.

Аналитическая обработка разного рода документов и формирование проблемно-ориентированных баз данных (ПОБД) в библиотеках и информационных органах может производиться для ИО и СИО конкретных коллективов ученых и разработчиков и/или для более широкого их использования через локальный или телекоммуникационный доступ, как в свободном, так и платном режиме. В последнем случае всем видам пользователей должно быть предоставлено в свободном доступе достаточно точное и представительное описание ресурса для его нахождения и принятия решения о приобретении, а сам ресурс предоставляется на платной основе. При этом важнейшими показателями эффективности работы служб, производящих аналитическую обработку документов и ПОБД для включения их описаний в электронные каталоги автоматизированных систем, являются точность и полнота поиска.

В современных условиях «точность поиска» должна избавить конечного пользователя от непроизводительных и трудоемких «ручных» процессов, связанных с просмотром и отбором пертинентных документов из числа найденных поисковой системой (релевантных и нерелевантных) документов, количество которых иногда исчисляется сотнями и даже тысячами. «Полнота поиска» призвана обеспечить нахождение всех документов, релевантных запросу для отбора пертинентных документов для пользователей. В то же время качество работы указанных критериев существенно зависит от разрешающей способности используемых при индексировании документов и запросов информационно-поисковых языков (ИПЯ), а также корректности их использования при обработке документов и запросов.

«Традиционными» и наиболее используемыми в большинстве российских библиотек являются ИПЯ классификационного типа, предназначенные для классификации и предметизации документов. В информационных органах, работающих в сфере научно-технической информации, предпочтение исторически отдается координатным (дескрипторным) языкам индексирования. Основные различия в этих ИПЯ и реализующих их словарных средствах (с одной стороны классификационных таблицах, с другой — тезаурусов и ключевых слов) заключаются в следующем:

Классификационные языки, являющиеся «предкоординатными», оперируют терминами, составляющими «рубрики» или «подрубрики», т. е заголовками, описывающими некоторые области или подобласти, в соответствующей области знания. Дескрипторные языки, являющиеся «(пост)координатными», ориентированы на точное указание объекта(ов) описания или поиска, в том числе — на фактографические сведения;

Статьи в рубрикаторах, описывающих наименование рубрик и подрубрик любого уровня, могут составлять сложные словосочетания и предложения. Лексические единицы дескрипторных языков строятся на коротких словосочетаниях (1-3 слова). При необходимости построения составных словосочетаний (например, в поисковых предписаниях) используются логические операторы, задающие разные виды отношений между лексическими единицами [1, 2, 3, 4, 5].

Таким образом, следует подчеркнуть, что вербальное индексирование позволяет описать разные аспекты содержания документов и БД с большей разрешающей способностью и более точно, чем классификационные ИПЯ. Исходя из изложенного, следует подчеркнуть важность использования для удовлетворения информационных потребностей ученых и разработчиков вербального (дескрипторного) индексирования, которое, кстати сказать, лежит в основе работы всех ИПС. При этом мы ни в коей мере не отвергаем необходимость классификационного индексирования, которое имеет важное значение в первую очередь для обеспечения различных внутрисистемных процессов (в библиотеке, группе библиотек, консорциуме и т. п.).

Традиционными процессами аналитической обработки документов и формирования локальных и распределенных электронных каталогов (ЭК) в библиотеках и информационных органах является составление библиографических записей (БЗ). Напомним, что в соответствии с ГОСТ 7.1-2003 БЗ включает библиографическое описание («совокупность библиографических сведений о документе, приведенных по определенным правилам, устанавливающим наполнение и порядок следования областей и элементов, и предназначенных для идентификации и общей характеристики документа»), заголовок библиографической записи (в соответствии с ГОСТ 7.1-2003), термины индексирования, аннотацию или реферат (см. ГОСТ 7.9-95), шифры хранения документа, справки о добавочных библиографических записях, дату завершения обработки документа, сведения служебного характера. Реализация указанных требований предусмотрена действующими коммуникативными форматами серии MARC, предусматривающими в составе БЗ соответствующих полей данных, в частности в UNIMARC`е и RUSMARC`е для предметных рубрик — 606 и индексов классификации — 675, 676, 679 и 686; для ключевых слов и персоналий — 610 и 600; для аннотаций и рефератов — 330 и т. п. [2, 6].

С большим сожалением следует отметить, что указанные требования выполняются в библиотеках не в полном объеме и далеко не всегда в удовлетворительном качестве:

Несмотря на то, что сотрудники подавляющего числа библиотек по данным анкетирования, проведенного среди участников АРБИКОНа в 2004 г. (более 150 библиотек разной ведомственной принадлежности) признают важность индексирования документов ключевыми словами (КС), реально используется этот способ аналитической обработки документов далеко не во всех библиотеках;

Качество индексирования КС по полноте и корректности отражения важных аспектов содержания документов также в большинстве случаев оставляет желать лучшего. В значительной степени это связано со следующими обстоятельствами: 1) Приверженностью сотрудников библиотек (особенно федеральных и вузовских) к традиционным способам и средствам индексирования (имеются ввиду «систематизация» и «предметизация») и в том числе — частая путаница принципов «вербального индексирования» с «предметизацией» (подробно об этом нами говорилось в докладах на международных конференциях КРЫМ 2004 и LIBCOM 2004 [7, 8]); 2) отсутствием согласованной методики составления КС (требования ГОСТ 7.66—92. «Индексирование документов.3) Отсутствием общепринятой нормативной терминологической базы для вербального индексирования, в частности тезаурусов и «авторитетных записей». Общие требования к координатному индексированию» явно недостаточно) — существуют разные методики, которые в ряде важных положений существенно противоречат друг другу [3, 7, 8];

Так же недостаточной является аналитическая обработка документов во многих библиотеках в части описания содержания документов для пользователей, поскольку в большинстве БЗ отсутствуют аннотации (не говоря уж о рефератах).

Очевидно поэтому, что гигантский объем литературы и содержащейся в ней информации остается не востребованным для пользователей как локальных, так и распределенных библиотечно-информационных систем. Особенно для тех из них, которые профессионально заняты в разных областях научных исследований и разработок.

Еще более серьезные требования к аналитической обработке документов и данных ставятся для их эффективного использования в режиме удаленного доступа. К ним относятся:

Требования обеспечения информационной совместимости в первую очередь для электронных каталогов и составляющих их БЗ это корректное использование коммуникативных (обменных) форматов и, в первую очередь — RUSMARC. А поскольку большинство внутренних форматов ПО АБИС ориентировано на форматы UNIMARC и UMARC 21 (USMARC), к тому же неоднозначно трактуемые, возникают проблемы совместного использования библиотечно-информационных ресурсов как библиотеками, так и их пользователями.

Описания отдельных полнотекстовых ресурсов, предоставляемых электронными библиотеками в Интернет или Интранет, созданными на базе библиотек и информационных органов, все более настоятельно требуют использования специальных систем метаданных и коммуникативных форматов, которые отражают виды этих ресурсов, характер их принадлежности, использования и т. п.

Термин «метаданные» принято толковать, как «данные о данных», однако его значение распространяется помимо описания состава данных, их структуры (формата) представления, места хранения и других признаков описания также на поддерживающие их информационные системы, технологии, пользователей, методы доступа и т. д. Особенно широко этот термин стал использоваться в последние годы в связи с развитием электронных библиотек, поскольку метаданные стали важнейшим средством обеспечения навигации, поиска и возможности информационного обмена в Интернете. Однако до настоящего времени значение этого термина до конца четко не определено. Наиболее размыты границы между метаданными и коммуникативными (обменными) форматами.

Существуют различные категории метаданных, например, описательные метаданные (в том числе библиографические); метаданные о структурах и форматах; административные метаданные, содержащие данные для управления доступом; идентификационные метаданные, которые однозначно идентифицируют объекты внешнего мира и т. п. Помимо сказанного, метаданные подразделяются на те, которые предназначены для автоматического решения определенного класса задач — машиночитаемые метаданные и тех задач, которые решаются с участием человека — человекочитаемые метаданные.

Наибольшее развитие в мировой информационной практике и России получила система Dublin Core, DC — Дублинское ядро (ДЯ), полное наименование системы: «Метаданные Дублинского ядра для простого открытия ресурса». Разработка ведется с 1995 г. рабочей группой с одноименным названием[1]. Ею предложена простая структура описания документов, которая, по мнению разработчиков, должна заменить сложные системы существующей каталогизации документов. Она предназначена для записи базовых структурных значений описания документов на языках разметки HTML и XML. Их состав включает в себя пятнадцать элементов, семантика которых была совместно определена международными группами профессионалов в

[1] Головные организации: **OCLC** (Online Computer Library Center) и **IETF** (Internet Engineering Task Force).

области библиотечного дела, вычислительной техники, кодирования текстов, специалистов музейного дела и других смежных областей наук.

В декабре 2000 г. в Лондоне на очередной выставке Online Information представители США, Англии, Франции, Германии и Японии называли DC наиболее перспективным стандартом метаданных для описания электронных ресурсов. Ряд национальных систем (например, Австралия, Швеция) уже объявили о принятии DC в качестве национального стандарта. В настоящее время ведется разработка версии DC 2. О. Рабочая группа Dublin Core работает в контакте с разработчиками RDF (структурная модель для выражения синтаксиса обмена метаданными, разработанная консорциумом W3C) [9, 10].

Помимо ДЯ существуют и вводятся в действия многие другие метаданные и коммуникативные форматы для представления в Интернете и обмена разнородными видами документов и данных и в частности:

CDIF* [CASE Data Interchange Format] — система стандартов, разрабатываемая и развиваемая организациями-членами Ассоциации EIA (Electronics Industries Standard). Их общая цель: стандартизация представления и обмена метаданными, описывающими различные информационные ресурсы, которые были созданы и поддерживаются с использованием различных технологий. Стандарты CDIF открывают возможности для повторного использования ресурсов метаданных в информационных системах для решения разнородных задач, в том числе для интеграции информационных ресурсов, полученных из различных источников. В настоящее время эти стандарты имеют статус внутренних для Ассоциации EIA, однако для придания им международного статуса они переданы на рассмотрение в ISO.

CSDGM (Content Standards for Digital Geospatial Metadata) — стандарт, разработанный Федеральным комитетом США по географической информации FGDC (US Federal Geographic Data Committee), предназначен для обеспечения обмена документами и данными о географическом пространстве. Он устанавливает имена элементов данных и их групп, используемых при обмене информационными ресурсами по данной тематике, а также сведения о значениях, которые должны присваиваться элементам данных разного рода.

Global Map (Specification for a data descriptive file for information interchange) — «Спецификация описательного файла цифровых географических данных для информационного обмена» представляет собой транспортный протокол OSI для структурированного обмена географическими данными. Разработана Международным координационным комитетом глобального картографирования — ISCGM (International Steering Committee for Global Mapping). Global Map позволяет создавать карты с разрешением в один километр, что эквивалентно обычному масштабу карты 1:1.000.000. Спецификация предусматривает восемь видов («уровней») географических данных: границы, перевозки (транспорт), дренаж, населенные пункты, возвышенности, растительность, почва и использование земель. Карты создаются в сотрудничестве с национальными картографическими организациями. В проекте участвуют 83 страны и региона, а более тридцати рассматривают такую возможность. Проектом охвачено 60% поверхности Земли.

HL7 (Health Level Seven) — «Здоровье уровня семь»: стандарт метаданных для обмена информацией в области здравоохранения. Его разработчиком является рабочая группа с одноименным названием при ANSI. Стандарт HL7 формализует интерфейсы между различными системами, обменивающимися сведениями о пациентах, включая данные анализов, назначений, результатов, оплаты и пр. Версия 2.3 включает также возможности для обмена информацией об уходе за пациентом, медицинских записях и автоматизированных инструментах. Все данные представлены знаками из выбранного набора (по умолчанию — ASCII). Версия 3.0 использует формализованную методику составления сообщений, описанную в HL7. Стандарт широко используется в госпиталях США, а также в Австралии, Германии, Японии, Голландии и Новой Зеландии. Он также является основой стандарта ISO 17113 (Method for Development of Messages).

ISAD (International Standard Archival Description) — «Международный стандарт архивного описания» содержит общие правила описания архивных документов. Разработан ICA (International Council on Archives). Вторая редакция документа одобрена в 1999 г. Он содержит правила записи 26-ти элементов описания единиц хранения, которые, как предполагается, могут использоваться в любых архивах. Каждое прави

ISO 11179 (Specification and Standardization of Data Elements) — «Спецификация и стандартизация элементов данных»: стандарт описания элементов данных в базах данных и документах. Разработан ISO/IEC JTC1/SC32 (публиковался по частям с 1994 по 2000 гг., последняя редакция опубликована в 2001 г.). Стандарт определяет базовые аспекты состава элемента данных (включая и метаданные) и предназначен для использования, как человеком, так и машиной, однако он не затрагивает проблем физического представления данных в виде последовательности битов на машинном уровне.

LOM (Learning Object Metadata) — «Метаданные учебного объекта«: стандарт, разработанный под эгидой IEEE Рабочей группой Компьютерного сообщества стандартизации — CSSAB (Computer Society Standards Activity Board) и Комитетом по стандартизации учебных технологий — LTSC (Learning Technology Standards Committee), для описания учебных ресурсов. Цель стандарта: облегчить поиск, рассмотрение и совместное использование учебных объектов учителями, инструкторами или автоматическими процессами в ходе выполнения учебных программ, а также обеспечить создание каталогов и хранилищ. Он предлагает

базовую схему, которая может использоваться для создания практических разработок. LOM является составной частью стандарта SCORM. Последняя спецификация стандарт IEEE LOM 1484.12 опубликована в июле 2002 г. Стандарт LOM 1484.12 является составным. Его части связанные с ISO 11404 (1484.12.2), XML (1484.12.3) и RDF (1484.12.4) находится на стадии рассмотрения.

OAIS (Reference Model for an Open Archival Information System) — «Образцовая модель для открытых архивных информационных систем»: модель метаданных, разработанная в 2002 г. Консультативным комитетом по космическим информационным системам CCSDS (Consultative Committee for Space Data Systems) и ISO TC20/SC13 для архивирования данных, связанных с космосом. Информационный блок OAIS содержит два вида данных: собственно контент (документы, базы данных и т. п.) и описание хранения PDI (Preservation Description Information).

PRISM (Publishing Requirements for Industry Standard Metadata) — «Требования к публикации для индустриального стандарта метаданных»: стандарт на метаданные, разработанный в 1999 г. некоммерческой организацией IDEAlliance PRISM Working Group. Его назначение: представление контента и описание формата, повторного и многоцелевого использования прав и ограничений на электронные ресурсы. PRISM разработан для использования в Интернете. Стандарт поддерживает ряд приложений, не содержит ограничений на формат данных описываемых ресурсов и построен на синтаксисе XML. Стандарт формулирует общие требования для обмена и хранения контента и метаданных (в виде коллекции элементов, описывающих контент), а также представляет набор контролируемых словарей, содержащих исчерпывающий перечень необходимых статей.

RDF* (Resource Definition Framework) — структурная модель для выражения синтаксиса обмена метаданными, разработанная консорциумом W3C. Последняя версия — RDF-Primer (см. <http://www.w3.org/TR/rdf-primer/>) рекомендована к использованию в феврале 2004 г. Для описания схемы метаданных и для обмена данными между различными вычислительными системами используется язык XML. RDF предлагает базовую систему типов, предназначенную для представление как данных, так и метаданных: «объект-атрибут-значение». Структурная модель состоит из «Ресурсов», «Типов свойства» и «Значений». Предоставляется возможность производить связь метаданных с различными информационными ресурсами и обмениваться метаданными между различными системами, которые их используют.

Z39.87 (Data Dictionary — Technical Metadata for Digital Still Images) — «Словарь данных: технические метаданные для неподвижных цифровых изображений»: проект стандарта (находится на стадии утверждения) разработан Организацией по национальным информационным стандартам США (NISO) и AIIM International в 2002 г. Он содержит полный список элементов технических терминов, необходимых для описания и управления техническим качеством цифровых неподвижных изображений (в том числе для поддержания их качества и обработки). Технические метаданные связываются с определенными атрибутами («якорями») качества изображения, которые могут быть объективно оценены: разрешение, тон, цвет, размер и т. п. [10, 11].

Действующие в Интернете и Рунете поисковые системы (Yandex, Rambler, Google, Yahoo!, Aport и др.) производят систематическое сканирование информационной среды Интернета, автоиндексирование и анализ выставленных ресурсов и обеспечивают пользователям нахождение нужных для них документов и данных. В указанном плане востребованность ресурсов и связанный с этим профессиональный (а может быть и коммерческий) успех организаций-создателей и владельцев этих ресурсов находятся в прямой зависимости от требований изложенных выше.

Становится все более очевидным, что отдельным даже очень крупным библиотекам и информационным органам в одиночку не возможно удовлетворительно решить указанные задачи. С этой целью и создаются корпоративные объединения библиотек. В частности в рамках Ассоциации региональных библиотечных консорциумов (АРБИКОН) определены обязательные для всех участников «профили» для БЗ в формате RUSMARC, ведутся работы по созданию единой методики вербального индексирования документов, рассматривается вопрос об участии совместно с национальными библиотеками в работе по созданию авторитетных записей, поставлен вопрос о начале рассмотрения вопросов, связанных с использованием метаданных и т. п.

Литература

1. Жданова Г. С., Колобродова Е. С., Полушкин В. А., Черный А. И. Словарь терминов по информатике / под ред. докт. Техн. наук, проф. А. И. Михайлова. — М.: Наука, 1971. — 359 с.
2. ГОСТ 7.1. —2003. Межгосударственный стандарт. Библиографическая запись. Библиографическое описание. Общие требования и правила составления. Взамен ГОСТ 7.1. —84, ГОСТ 7.16. —79 (нотные издания), ГОСТ 7.18. —81 (картографические издания), ГОСТ 7.34. —81 (изоиздания), ГОСТ 7.40. —82 (аудиовизуальные издания). — Введ. с 01.07.2004 г.
3. ГОСТ 7.9—95 (ИСО 21 4-76). Реферат и аннотация. Общие требования. — Взамен ГОСТ 7.9—77. — Введ.01.07.97 г.

4. ГОСТ 7.66—92. (ИСО 5963—85). Индексирование документов. Общие требования к координатному индексированию. — Введ.01.01.93 г.
5. ГОСТ 7.25—2001. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав, форма. Взамен ГОСТ 7.25—80. — Введ.01.07.2002 г.
6. ГОСТ 7.82—2001. Межгосударственный стандарт. Библиографическая запись. Библиографическое описание электронных ресурсов. Общие требования и правила составления. — Введ. (впервые) 01.07.2002 г.
7. Воройский Ф. С., Острая С. А. Информационные ресурсы АРБИКОНа — пора подумать об их качестве // «Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества»: Материалы конф. «КРЫМ-2004». М.:ГПНТБ России, 2004 г. http://www.gpntb.ru/win/inter-events/crimea2004/disk/chapter2.html#section11/.
8. Воройский Ф. С. Основные принципы обеспечения информационного поиска в корпоративных электронных каталогах // «Информационные технологии, компьютерные системы и издательская продукция для библиотек»: доклады и тезисы докладов МК «LIBCOM-2004». — М.: ГПНТБ России, 2004. — С.56-59.
9. Arms William Y. Digital Libraries. — Cambridge, Massachusetts, London, England.: The MIT Press, 2000. — 287p.
10. Хохлов Ю. Е. Обзор форматов метаданных. Российские электронные библиотеки. [Электронный ресурс] / Ю. Е. Хохлов, С. А. Арнаутов. — Режим доступа: http://www.elbib.ru/index. phtml?page=elbib/rus/methodology/mdrev. — Заглавие с экрана.
11. Воройский Ф. С. Информатика. Энциклопедический словарь-справочник. Введение в современные информационные технологии в терминах и фактах. Издание четвертое переработанное и дополненное. / Ф. С. Воройский. — М.: Физматлит, 2006. — 945 с. (в печати).