УДК 025.17:004]:004.738.5

И. М. Зацман

Электронные библиотеки научных документов в Интернет: структуризация, формальное описание и поиск невербальной информации

Рассматривается проблема структуризации, формального описания и поиска научных полнотекстовых документов в политематических электронных библиотеках при условии, что электронная форма представления документов включает вербальные и структурно-графические (невербальные) компоненты. Показано, что создание в сети Интернет политематических электронных библиотек, интегрирующих большие объемы научной информации по разным областям знаний, приводит к возникновению новых задач поиска документов по их вербальным и структурно-графическим компонентам.

ВВЕДЕНИЕ

Создание политематических электронных библиотек полнотекстовых научных документов позволяет упростить и существенно ускорить доступ к первоисточникам научной информации. Необходимость концентрации усилий на решении проблемы доступа к научным первоисточникам в электронной форме была зафиксирована в итоговых документах Первой сессии Международного совета ЮНЕСКО по глобальным научным коммуникациям в сентябре 1995 г. В них говорится о необходимости широкого использования электронной формы представления результатов современных научных исследований и создания научных документов изначально в электронном виде, формирования ретроспективных электронных фондов на основе уже существующих в бумажной форме традиционных научных трудов и изданий, а также обеспечения к ним доступа через Интернет [1].

Одной из попыток решения проблемы доступа к научным документам в электронной форме является исследовательская программа Digital Library Initiative, финансируемая NSF, ARPA и NASA, основной целью которой является создание, развитие и эффективное использование цифровых библиотек. Первый этап программы охватывает период: сентябрь 1994 г. — август 1998 г. В рамках программы Digital Library Initiative термин *цифровые библиотеки* используется для обозначения хранилищ, рассчитанных на более широкий (по сравнению с научными документами) перечень видов документов. Однако часть проектов этой программы относится именно к электронным библиотекам научных документов. Для участия в программе Digital Library Initiative был объявлен конкурс. По результатам конкурса для финансирования отобраны проекты, которые заявили следующие университеты: Иллинойский, Стэнфордский, Мичиганский, Бэркли, Карнеги—Мэллон и Санта-Барбара (Калифорния). Всего было отобрано для выполнения шесть проектов (из 73 заявленных) [2]. В России с 1998 г. ряд министерств и ведомств начали финансирование на конкурсной основе Межведомственной программы "Российские электронные библиотеки" [3].

Данная статья посвящена политематическим электронным библиотекам, которые, кроме хранения полнотекстовых предметно-ориентированных научных документов, относящихся к разным областям знаний (математика, механика, физика, астрономия, химия, науки о жизни, науки о Земле), могут включать и политематические научные документы, относящиеся одновременно к нескольким областям знаний. Интеграция больших объемов документов, относящихся к разным областям знаний, стала основной причиной возникновения более сложных поисковых задач, которые не могут быть решены с помощью традиционных методов индексирования и поиска вербальной информации [4].

Политематические электронные библиотеки, интегрирующие большие объемы электронных форм традиционных научных документов (статей, монографий, диссертаций, научных отчетов, трудов конференций, семинаров) по различным областям знаний, иногда рассматривают как автоматизированный аналог традиционных библиотек. Однако переход от бумажной к электронной форме представления влечет за собой следующие кардинальные изменения:

хранение научного первоисточника в цифровой форме;

возможность телекоммуникационного доступа;

возможность тиражирования научных первоисточников помимо традиционных издательских и полиграфических процессов.

Анализ последствий перечисленных изменений можно найти в работе [5].

По функциональным возможностям электронные библиотеки полнотекстовых документов качественно отличаются от традиционных библиотек. В первую очередь, кроме библиографического поиска, который обеспечивается библиотечными каталогами (бумажными или электронными), в элек-

тронных библиотеках можно сочетать поиск по реквизитам документов и поиск по полным текстам. Пользователи электронных библиотек могут получать полные или выборочные копии найденных документов в электронной и/или бумажной формах. При этом документы в электронных библиотеках могут храниться одновременно в текстовой и факсимильной формах [6]. Кроме традиционных научных документов, в электронных библиотеках могут храниться научные документы, изначально создаваемые в электронном виде.

Ниже предпринята попытка сформулировать проблему структуризации, формального описания и поиска полнотекстовых научных документов в политематических электронных библиотеках. При поиске в электронных библиотеках предлагается не ограничиваться только вербальными компонентами полнотекстовых научных документов. Рассматриваются следующие аспекты проблемы структуризации, формального описания и поиска научных документов в политематических электронных библиотеках:

вербальные и структурно-графические (невербальные) компоненты научных документов (особенности представления);

электронные формы кодирования компонентов традиционных научных документов;

моделирование научных документов, относящихся к разным областям знаний;

создание информационно-поисковых систем для электронных библиотек полнотекстовых научных документов.

ВЕРБАЛЬНЫЕ И СТРУКТУРНО-ГРАФИЧЕСКИЕ КОМПОНЕНТЫ НАУЧНЫХ ДОКУМЕНТОВ

Под вербальными компонентами документа здесь понимаются линейные алфавитно-цифровые фрагменты названия документа, аннотации, разделов, глав и параграфов документа; под структурно-графическими (невербальными) компонентами — математические и структурные химические формулы, таблицы, графики, схемы, рисунки, фотографии, карты; под полнотекстовым документом — совокупность всех вербальных и структурно-графических его компонентов; под полнотекстовым поиском понимается поиск только по вербальным компонентам документов.

Принципиально нового качества поиска полнотекстовых научных документов в электронных библиотеках можно достичь, дополнив реквизитный и полнотекстовый поиск по вербальным компонентам поиском научных документов по структурно-графическим компонентам. Важной особенностью поиска научных документов по структурно-графическим компонентам является отсутствие языкового барьера, являющегося неотъемлемой чертой поиска по вербальным компонентам.

Новое качество, которое получает пользователь при поиске по структурно-графическим компонентам научных документов, является следствием того, что научные знания, отраженные в структурно-графических компонентах, могут не получить адекватного описания в вербальных компонентах того же самого научного документа. С точки зрения содержания научных документов, для пользователей электронной библиотеки структурно-графические компоненты научных документов могут быть даже более ценными, чем их вербальные составляющие. Поэтому поиск по структурно-графическим компонентам может представлять бóльшую ценность, чем поиск по вербальным компонентам научных документов.

Наиболее ярким примером такой ситуации являются научные документы по химии, в которых структурная химическая информация является для пользователя семантически не менее (а скорее всего и более) значимой, чем вербальные компоненты документов. Поэтому и получили большое распространение такие документальные химические базы данных, в которых поиск можно вести по структурам химических соединений. Кроме документальных баз данных широко используются базы структурных данных по химии [7, 8].

Кроме химии, предпринимаются попытки разработать методы и алгоритмы формального описания и кодирования структурно-графических компонентов и в других областях знаний. В картографических информационных системах, когда известны классы знаковых (графических) элементов, цельных конфигураций, которые предстоит выделять и анализировать в картах, как правило, заранее разрабатывается словарь конструктивных элементов и грамматические правила построения описаний [9]. Этот подход может служить основой для разработки методов и алгоритмов поиска геологических и географических научных документов по картографической информации [10].

Отметим, что существующие в настоящее время методы и алгоритмы электронного представления картографической информации, используемые в геоинформационных системах, как правило, не пригодны для поиска полнотекстовых геологических и географических научных документов в электронных библиотеках по их картографическим компонентам (для тех документов, в которых они присутствуют). Однако существующие методы и алгоритмы электронного представления могут быть использованы при отображении картографической информации в геологических и географических документах, найденных по вербальным компонентам.

Из структурно-графических компонентов, поиск по которым реализован в политематических электронных библиотеках, необходимо отметить таблицы, которые содержат только вербальную информацию [11].

В настоящее время отсутствуют постановка и решение проблемы структуризации и формального описания полнотекстовых научных документов разных областей знаний в электронных библиотеках с целью организации поиска одновременно по вербальным и структурно-графическим компонентам. Например, выполнение в политематических электронных библиотеках запроса на поиск, состоящего из подзапросов:

на поиск структурной химической информации,

на поиск по картографическим компонентам,

на поиск по геохимическим диаграммам

и одновременно полнотекстового подзапроса с учетом контекстной близости перечисленных компонентов, является в настоящее время нерешенной задачей.

1-7-5485

Иногда при создании политематических электронных библиотек с помощью традиционных информационных технологий, в качестве единой электронной формы представления структурно-графических компонентов научных документов (или всего документа в целом) используется факсимильное представление, которое дает возможность отображать структурно-графические компоненты (или полноразмерные страницы документа), но не позволяет организовать по ним поиск.

Поиск полнотекстовых научных документов в политематических электронных библиотеках по вербальным и структурно-графическим компонентам, включая обеспечение поиска через Интернет, требует создания новых информационных технологий структуризации, учитывающих специфику научных документов разных областей знаний, и формального описания встречающихся в них структурно-графических компонентов с целью расширения области поиска за счет невербальной информации документов.

КОДИРОВАНИЕ КОМПОНЕНТОВ НАУЧНЫХ ДОКУМЕНТОВ

Ограничимся рассмотрением электронных форм представления только традиционных научных документов (статей, монографий, диссертаций, научных отчетов, трудов конференций, семинаров), которые в электронной библиотеке, как правило, образуют совокупность декларативных знаний. Научные документы, изначально создаваемые в электронной форме, в общем случае могут представлять собой сочетание декларативных и процедурных знаний [12].

Решение задачи электронного представления и кодирования вербальных и структурно-графических компонентов традиционных научных документов зависит от тех возможностей компьютерной обработки и хранения данных, которые во многом определяются архитектурой вычислительных систем, обеспечивающих функционирование электронных библиотек.

В современных вычислительных системах, как правило, используется побайтовая организация памяти (при этом код символа может быть однобайтовым, двухбайтовым или многобайтовым). Чтобы реализовать возможности компьютерной обработки и хранения данных, научные документы на системно-аппаратном уровне должны быть представлены в виде линейных последовательностей байтов.

Рассмотрим следующие электронные формы представления данных в виде линейных последовательностей байтов:

алфавитно-цифровая (когда линейной последовательности алфавитно-цифровых символов в документе соответствует линейная последовательность байтов);

растровая (когда линейной последовательности точек изображения соответствует линейная последовательность байтов);

интерпретируемая символьно-кодированная, которая используется для представления тех структурно-графических компонентов, для которых уже разработаны методы линейного кодирования и интерпретации закодированных компонентов.

В настоящее время имеются информационные технологии подготовки научных документов и формирования полнотекстовых баз данных, включая алфавитно-цифровую, растровую и интерпретируемую символьно-кодированную формы представления [13, 14].

Алфавитно-цифровая форма представления применяется для вербальных компонентов документов — линейных последовательностей символов. Растровая форма представления используется для рисунков, фотографий, графиков, карт, чертежей и схем, иногда — для факсимильного представления полноразмерных страниц документов. В этом случае каждая страница отображается как растровое изображение.

Большинство структурно-графических компонентов традиционных научных документов в настоящее время могут быть представлены только в растровой форме; отдельные структурно-графические компоненты — в интерпретируемой символьно-кодированной форме (такое представление компонента будем называть интерпретируемым символьно-кодированным объектом). Минимально необходимый набор электронных форм декларативного представления данных в вычислительных системах для хранения и обработки вербальных и структурно-графических компонентов традиционных научных документов включает алфавитно-цифровую, растровую и интерпретируемую символьно-кодированную формы представления. Отметим, что для поиска научных документов по их структурно-графическим компонентам пригодна не любая символьно-кодированная форма представления невербальных компонентов.

Традиционные информационные технологии создания, развития и использования политематических электронных библиотек делают, как правило, возможным поиск полнотекстовых научных документов только по их вербальным компонентам [15]. При этом структурно-графические компоненты научного документа могут храниться в политематических электронных библиотеках как дополняющие вербальные компоненты иллюстрации, интерпретация которых, с точки зрения поиска документов, невозможна. Однако, если обеспечивается хранение, то имеется возможность отображения и печати иллюстраций научных документов, найденных в электронных библиотеках по вербальным компонентам.

Если говорить о поиске по структурно-графическим компонентам, то такой поиск в настоящее время реализован в отдельных предметно-ориентированных электронных библиотеках и базах данных. В первую очередь, необходимо отметить базы данных по химии, в которых возможен поиск по структурной информации. При обработке научных документов по химии на стадии подготовки данных используется символьное кодирование структурных формул. Алгоритмы кодирования строятся таким образом, чтобы была возможность идентифицировать структурную информацию на основе соответствующих ей интерпретируемых символьно-кодированных объектов, а также осуществить поиск документов по структурной химической формуле или ее фрагментам [16]. Методы и алгоритмы линейного кодирования структурной химической информации, которые применяются для создания предметно- или проблемно-ориентированных баз дан-

ных по химии, могут также использоваться при создании политематических электронных библиотек, но только для данного вида структурно-графических компонентов (структурной химической информации). При этом эти методы и алгоритмы, как правило, не применимы для других видов структурно-графических компонентов научных документов.

Для создания, развития и эффективного использования политематических электронных библиотек, в которых возможен поиск по вербальным и структурно-графическим компонентам научных документов, относящихся к разным областям знаний, необходимо разработать новые методы, алгоритмы и информационные технологии структуризации полнотекстовых научных документов и формального описания их компонентов в рамках некоторой модели научного документа (подразумевается, что для формального описания используются искусственные, а не естественные языки).

МОДЕЛИРОВАНИЕ НАУЧНЫХ ДОКУМЕНТОВ

Для поиска научных документов в электронных библиотеках разработка методов и алгоритмов структуризации полнотекстовых научных документов и формального описания их компонентов становится ключевой задачей. Для многих видов структурно-графических компонентов задача их формального описания даже и не полностью формализована.

Разработку методов и алгоритмов структуризации полнотекстовых научных документов и формального описания с целью распознавания и поиска их компонентов (включая структурно-графические компоненты) будем называть задачей построения логико-семантической модели научных документов (в соответствии с точкой зрения Ю. И. Шемакина и А. А. Романова, которые пишут о закономерности перерастания логико-лингвистических моделей в логико-семантические [17]). Построение логико-семантической модели научных документов для проектирования на ее основе электронных библиотек можно рассматривать как один из подходов к проблеме интеграции научных информационных ресурсов.

Задача построения логико-семантической модели научных документов, интегрирующей методы и алгоритмы формального описания математических и структурных химических формул, таблиц, графиков, схем, карт, рисунков и фотографий, в настоящее время не решена. В приведенный перечень включены изображения (рисунки и фотографии), задачи распознавания и идентификации которых представляют отдельную проблему, имеющую отношение не только к научным документам.

В настоящее время разработаны методы и алгоритмы формального описания только отдельных видов структурно-графических компонентов (например, уже упоминавшихся структурных химических формул и картографической информации). При этом не всегда на основе алгоритмов формального описания отдельных видов структурно-графических компонентов удается достаточно просто создать алгоритмы индексирования этих компонентов и поиска по ним.

Задача построения логико-семантической модели научных документов ориентирована, в первую очередь, на задачи поиска, что является принципиальным ее отличием от других задач моделирования документов (например, построения объектной модели документа, ориентированной на решение задачи навигации по документу [18]). Остановимся на целевой ориентации объектной модели документа, предложенной консорциумом World Wide Web (W3C), более подробно.

В октябре 1997 г. консорциум W3C предложил проект спецификации объектной модели документа. В основу проекта спецификации объектной модели документа положена иерархическая модель данных. Эта спецификация определяет объектно-ориентированный интерфейс, обеспечивающий доступ к Web-страницам, HTML- и XML-документам (т. е. речь идет не только о научных документах). На верхнем уровне иерархии находится сам документ. На следующем размещаются такие его элементы, как заголовки, параграфы, абзацы и комментарии. Атрибуты и тексты каждого из этих элементов располагаются уровнем ниже. Модель документа описывается на IDL (Interface Definition Language — язык описания интерфейсов) интероперабельной архитектуры CORBA (Общая архитектура брокера объектных заявок — Common Object Request Broker Architecture). Однако проект спецификации не требует, чтобы каждая реализация спецификации объектной модели документа была основана на CORBA [18].

В проекте спецификации объектной модели документа (Document Object Model — DOM), который предложен консорциумом W3C, предполагается, что DOM-приложение получает уже готовую ссылку на документ. Если поиск документов в Интернет надо вести только по вербальным компонентам и/или реквизитам (атрибутам) хранимых документов, то создание методов и алгоритмов получения (вычисления) ссылок на документы не требует построения логико-семантических моделей. Необходимые ссылки на документы можно получить, например, с помощью поисковых серверов Интернет, алгоритмы индексирования которых основаны на логико-лингвистических моделях документов, не учитывающих содержание структурно-графических компонентов научных документов. После получения готовой ссылки на документ DOM-приложение использует объектную модель документа консорциума W3C для навигации по документу и, если необходимо, для выполнения операций над документом, его атрибутами и элементами (операции добавления, удаления, изменения атрибутов, элементов документа).

Таким образом, объектная модель документа определяет некоторый стандартный набор объектов для представления HTML- и XML-документов, методы и алгоритмы комбинирования этих объектов, а также интерфейс для доступа к ним и выполнения операций над ними. Однако объектно-ориентированный подход, в рамках которого консорциум W3C строит модель документа для навигации, может быть основой построения логико-семантической модели научных документов для обеспечения поиска в электронных библиотеках (наряду с обеспечением доступа к документам и выполнения операций над ними).

Проект спецификации объектной модели доку-

мента консорциума W3C основан на единственном способе организации объектов (иерархическом). Один из возможных подходов к построению логико-семантической модели научных документов состоит в применении нескольких способов организации объектов с учетом их вложенности. При использовании вложенности способ организации объектов на самом верхнем уровне модели научного документа будем называть базовым или опорным. При этом не предполагается ограничиваться только традиционным набором способов организации (иерархический, сетевой и реляционный).

Например, для представления в политематических электронных библиотеках геологических и географических научных документов полезным может оказаться пространственно-временной способ организации объектов [19]. Моделирование таких документов требует определения некоторого набора объектов для их представления, методов и алгоритмов комбинирования этих объектов. При этом необходимо иметь возможность оперировать тремя координатами и временем. Пространственно-временной способ организации объектов позволяет более точно отобразить содержание этих документов и представить в электронных библиотеках не только электронные формы традиционных научных документов (изначально созданных на бумаге), но и создаваемые изначально в электронном виде, включающих представление динамических процессов (например, изменение геологических структур во времени).

В качестве примера традиционного научного документа в бумажной форме, для электронного представления содержания которого наряду с другими способами организации объектов удобно было бы использовать пространственно-временной способ, рассмотрим Словарь географических названий форм подводного рельефа [20]. Каждая словарная статья этого документа в бумажном издании содержит три элемента: название формы подводного рельефа на русском языке, название в международной транскрипции и географические координаты. Если для электронного представления использовать иерархический способ организации объектов (как в проекте спецификации DOM), то все словарные статьи можно рассматривать как узлы одного уровня иерархии, что дает возможность структурировать электронное представление Словаря аналогично его бумажному изданию. При этом все узлы можно упорядочивать по любому из трех элементов словарной статьи (в бумажной форме словарные статьи упорядочены по названиям форм на русском языке и дополнены указателем названий в международной транскрипции).

Если для электронного представления данных Словаря применить пространственно-временной способ организации объектов, то можно включить в электронное представление Словаря дополнительную информацию о формах подводного рельефа, которая накоплена в результате исследований, но не вошла в существующий Словарь из-за естественной ограниченности традиционных изданий по объему и из-за их статичности при отображении научной информации. Принципиально новые возможности электронного представления научной информации, которые предоставляют пространственно-временной способ организации объектов, широко используются в геоинформационных системах.

Предлагаемый подход к построению логико-семантической модели научных документов, относящихся к разным областям знаний, состоит в сочетании нескольких способов организации объектов с учетом возможной их вложенности. При вложенности базовый (опорный) способ организации данных на самом верхнем уровне определяется в момент создания электронной формы представления научного документа. При этом возможны случаи, когда один документ имеет несколько электронных форм представления на основе различных базовых способов организации объектов. Важно отметить, что логико-семантическая модель научных документов и способы организации объектов вместе составляют внешнюю модель, от которой, в общем случае, не зависят модели хранения данных в электронной библиотеке.

Сочетания нескольких способов организации объектов с учетом их вложенности недостаточно для построения логико-семантической модели научных документов. Необходимо также определить и набор объектов для формального представления структурно-графических компонентов, которые могут быть использованы при поиске в политематических электронных библиотеках.

ИНФОРМАЦИОННО-ПОИСКОВЫЕ СИСТЕМЫ ДЛЯ ЭЛЕКТРОННЫХ БИБЛИОТЕК НАУЧНЫХ ДОКУМЕНТОВ

Объектно-ориентированный подход, в рамках которого консорциум W3C строит модель документа для решения задач навигации, в настоящее время широко используется при разработке объектно-ориентированных баз данных, систем управления этими базами данных (ОСУБД) и применения ОСУБД для приложений мультимедиа [21, 22, 23]. В объектно-ориентированных системах управления базами данных уже апробирован ряд подходов к представлению и хранению данных сложной структуры со сложными взаимосвязями, но общей (точнее, единой) модели данных для объектно-ориентированных БД в настоящее время нет и разработчики существующих ОСУБД используют, как правило, несовместимые между собой объектные модели данных [24]. Несовместимость объектных моделей является одним из основных препятствий (но не единственным) для их применения в качестве основы решения задачи моделирования научных документов. Кроме того, в настоящее время для целого ряда приложений с помощью существующих ОСУБД не удается решить задачу автоматического индексирования [21]. Однако, если задачу поддержки логико-семантической модели научных документов решать на другом (более высоком) уровне организации процесса поиска, то отдельные СУБД и ОСУБД могут быть использованы на более низком уровне для хранения в гетерогенных БД стандартных и специализированных типов данных.

Для реализации многоуровневого поиска в политематических электронных библиотеках научных документов, кроме построения их логико-семантической модели, необходимо разработать принципиально новые информационно-поисковые системы (ИПС). ИПС, спроектированные на основе логико-семантической модели, будем называть *объектно-ориентированными*, в отличие от традиционных

ИПС (обзор традиционных ИПС см., например, в [25]).

Для поиска подходов к построению объектно-ориентированных ИПС рассмотрим одно из проектных решений фирмы Sybase, использованное при создании Adaptive Server в рамках трехуровневой архитектуры Adaptive Component Architecture (клиент — промежуточный слой — сервер) [26]. Adaptive Server обеспечивает унифицированную работу с гетерогенными источниками и специализированными типами данных. За интеграцию различных источников и типов данных отвечает своеобразная программная шина Component Integration Layer. Эта шина обеспечивает совместную работу собственных продуктов фирмы Sybase непосредственно, а также делает возможным подключение СУБД иных производителей через специальные модули DirectConnect. Фирма Sybase, в отличие от основных конкурентов, не пытается создавать единый универсальный сервер для всех типов данных, а использует для специализированных типов данных отдельные компоненты, являющиеся продуктами независимых производителей и удовлетворяющие открытым стандартам. Для работы с географическими данными используется продукт SQS фирмы Vision, с мультимедиа-информацией — продукт 1View:Object Manager компании Network Image, с большими текстами — продукт Verity Text-Search фирмы Verity, с временными рядами — продукт FAME одноименной фирмы и с изображениями — продукт Virage Entertainment Insurance компании Virage. Таким образом, реализация трехуровневой архитектуры Adaptive Component Architecture базируется на компонентном подходе.

Развитие этого решения фирмы Sybase, направленное на применение единого монитора поиска, может служить основой проектирования объектно-ориентированных ИПС для политематических электронных библиотек научных документов. При применении ИПС с трехуровневой архитектурой пользователям электронных библиотек будет доступен только клиентский уровень, на котором обеспечивается построение запросов, работа со словарями ИПС и/или с тезаурусом, списками найденных документов, вербальными и структурно-графическими компонентами найденных документов, персональной информацией пользователя (включая сохраненные запросы). Пользователь при работе с ИПС может применить логико-семантическую модель в самом общем виде или ее частные случаи (подмодели), что позволит адаптировать поисковые возможности ИПС и набор разрешенных для включения в поисковые запросы объектов.

Например, ограничив область поиска только вербальными компонентами научных документов, пользователь сможет осуществлять традиционный полнотекстовый поиск. Добавив к области поиска компоненты со структурной химической информацией, пользователь уже сможет включить в запрос структурные объекты и реализовать с помощью одного запроса поиск по структурам и вербальным компонентам, определив при этом степень необходимой ему их контекстной близости.

Монитор поиска, к которому поступают поисковые запросы пользователя, может быть реализован на промежуточном уровне. В настоящее время такая реализация мониторов транзакций является стандартным решением целого ряда серийно выпускаемых продуктов. Реализация на этом уровне мониторов поиска документов по вербальным и структурно-графическим компонентам научных документов в настоящее время не осуществлена.

В ИПС с трехуровневой архитектурой, на серверном уровне реализуется хранение вербальных и структурно-графических компонентов научных документов, включая необходимую поддержку хранения в гетерогенных базах данных стандартных и специализированных типов данных. Трехуровневая архитектура ИПС позволяет также организовать согласованное функционирование нескольких СУБД, ориентированных на разные специализированные типы данных, что может быть реализовано совместно с распределенным хранением научных информационных ресурсов.

Реализации объектно-ориентированной ИПС, обеспечивающей поиск по вербальным и структурно-графическим компонентам, должно предшествовать построение логико-семантической модели научных документов, что само по себе представляет достаточно сложную задачу. Однако компонентный подход к построению объектно-ориентированной ИПС дает возможность решать задачу поиска научных документов по структурно-графическим компонентам поэтапно. На первом этапе может быть реализован поиск в электронных библиотеках только по тем структурно-графическим компонентам, для которых уже существуют методы и алгоритмы кодирования/интерпретации и индексирования. По мере создания соответствующих методов и алгоритмов для других компонентов, включая определение перечней объектов для их представления, будут развиваться логико-семантическая модель и поисковые возможности объектно-ориентированных ИПС, предназначенных для политематических электронных библиотек.

Компонентный подход к построению объектно-ориентированной ИПС с трехуровневой архитектурой предполагает существование универсальной для всех компонентов среды исполнения. Этот вопрос выходит за рамки статьи и в силу его важности заслуживает отдельного рассмотрения.

ЗАКЛЮЧЕНИЕ

Разработка новых информационных технологий создания, развития и эффективного использования политематических электронных библиотек, интегрирующих большие объемы научных документов по всем областям знаний с возможностью поиска по структурно-графическим компонентам документов, во многом зависит от построения логико-семантической модели и разработки объектно-ориентированных ИПС.

Спецификация объектной модели документа, предлагаемой в настоящее время консорциумом W3C, ориентирована, в первую очередь, на задачи навигации и выполнение операций над документом, его атрибутами и элементами (операции добавления, удаления, изменения атрибутов, элементов документа). Задача построения логико-семантической модели полнотекстовых научных документов имеет принципиально другую целевую ориентацию, а именно: поиск в электронных библиотеках по вербальным и структурно-графиче-

1-9-5485

ским компонентам научных документов. Построение логико-семантической модели полнотекстовых научных документов в виде открытого и развиваемого стандарта (по аналогии со спецификацией объектной модели документа консорциума W3C) решило бы проблему интеграции научных информационных ресурсов для мета-сети Интернет в рамках электронной мета-библиотеки следующего тысячелетия (в работе [12] подобная мета-библиотека называется World Digital Library System). Для работы с электронной мета-библиотекой потребуются объектно-ориентированные многоуровневые ИПС нового поколения.

В настоящее время можно говорить о существовании методов формального описания, алгоритмов кодирования/интерпретации и индексирования только для некоторых структурно-графических компонентов, ориентированных, как правило, на отдельные научные дисциплины. В этих случаях поиск в предметно-ориентированных базах данных по соответствующим компонентам успешно реализуется сегодня с помощью специализированных СУБД на стандартных средствах вычислительной техники, производительность которой позволяет обрабатывать запросы на поиск по этим структурно-графическим компонентам в приемлемое для пользователей время.

С точки зрения технических возможностей ЭВМ, интересно сравнить сегодняшнее положение с хранением и поиском полнотекстовых документов с ситуацией, описанной Дж. Солтоном более двадцати лет назад [27]. Он пишет: "Использование двойной технологии, включающей системы хранения как на магнитных лентах, так и на фотоносителях, порождает сложные проблемы, так как стандартное вычислительное оборудование не приспособлено для управления микрохранилищами. Однако экспериментально эти проблемы были решены. Так созданы пультовые устройства, которые могут поочередно принимать данные из памяти ЭВМ и микрозаписи из памяти на фотоносителях... " [27, с. 159]. Отметим, что методы поиска, описанные в книге Дж. Солтона, обеспечивали теоретическую основу для решения задачи полнотекстового поиска документов. Однако реализация этих методов на стандартных средствах вычислительной техники того времени была невозможна. Ограниченные возможности ЭВМ позволяли хранить только библиографические описания и рефераты, которые использовались для поиска с целью последующего обращения к микрохранилищу за фотоформами полнотекстовых документов. Затем рост технических возможностей ЭВМ позволил хранить в цифровом виде полные тексты документов и осуществлять полнотекстовый поиск документов.

Положение в области поиска научной информации, сложившееся сейчас, качественно отличается от ситуации, описанной Дж. Солтоном. В настоящее время успехи в разработке информационных технологий создания, развития и эффективного использования политематических электронных библиотек, возможно, будут определяться результатами современных фундаментальных исследований в области моделирования научных документов и поиска научной информации. А новые методы и алгоритмы поиска научных документов по их структурно-графическим компонентам смогут быть реализованы на стандартных средствах вычислительной техники, доступных на рубеже тысячелетий.

СПИСОК ЛИТЕРАТУРЫ

1. International Advisory Council on Global Scientific Communications (ACOSC) // Report of First Session.— Paris: UNESCO, 1995.

2. Schatz B., Chen H. Building Large-Scale Digital Library // Computer.— 1996 .— Vol. 29, № 5 .— P. 22–26.

3. Межведомственная программа "Российские электронные библиотеки" // Газета "Поиск". 1998 .— № 13(463), 21–27 марта.— С. 7.

4. Knoblock C. et al. The Role of AI in Digital Libraries // IEEE Expert.— 1995 .— Vol. 11, № 3 .— P. 8–13.

5. Wiederhold G. Digital Library, Value, and Productivity // Comm. of the ACM.— 1995 .— Vol. 38, № 4 .— P. 85–96.

6. Entlich R., Garson L., Lesk M. et al. Making a Digital Library: The Chemistry Online Retrieval Experiment // Comm. of the ACM.— 1995 .— Vol. 38, № 4 .— P. 54.

7. Derwent Databases // Derwent Scientific and Patent Information: Online Databases Catalogue. — London: Derwent Information Limited, 1996 .— P. 4–50.

8. Алфимов М. В., Авакян В. Г., Джигирханова А. В., Буторина Л. С. Фосфор-серо- и кремний содержащие соединения в базе структурных данных по химии // Информационные продукты, процессы и технологии: НТИ-95: Материалы конф. (Москва, 19–20 октября 1995 г.). Т. 1.— М.: ВИНИТИ, 1995 .— С. 117.

9. Лютый А. А. Язык карты: сущность, система, функция.— М.: ИГ АН СССР, 1988 .— С. 244.

10. Wilensky R. Toward Work-Centered Digital Information Services // Computer.— 1996 .— Vol. 29, № 5 .— P. 37–43.

11. Croft W. B. NSF Center for Intelligent Information Retrieval // Comm. of the ACM.— 1995 .— Vol. 38, № 4 .— P. 42–43.

12. Fox E. A., Akscyn R. M., Furuta R. K., Leggett J. J. Digital Library (Introduction) // Comm. of the ACM.— 1995 .— Vol. 38, № 4 .— P. 23–28.

13. Генин Б. Л. Развитие информационного обслуживания пользователей патентной информацией // Информационные продукты, процессы и технологии: НТИ-95: Материалы конф. (Москва, 19–20 октября 1995 г.). Т. 1 .— М.: ВИНИТИ, 1995 .— С. 56–59.

14. Зацман И. М. Полнотекстовые многотомные базы данных на CD-ROM (Технологические аспекты формирования) // Информационные продукты, процессы и технологии: НТИ-95: Материалы конф. (Москва, 19–20 октября 1995 г.). Т. 2 .— М.: ВИНИТИ, 1996 .— С. 19–31.

15. Schatz B., Cole T. W., Hardin J. B. et al. Federating Diverse Collection of Scientific Literature // Computer.— 1996 .— Vol. 29, № 5 .— P. 28–35.

16. Бессонов Ю. Е., Красотченко В. В. Алгоритм идентификации химических соединений // Информационные продукты, процессы и технологии: НТИ-95: Материалы конф. (Москва, 19–20 октября 1995 г.). Т. 1 .— М.: ВИНИТИ, 1995 .— С. 119.

17. Шемакин Ю. И., Романов А. А. Компьютерная семантика.— М.: НОЦ "Школа Китайгородской", 1995 .— 344 с.

18. Салливан И. DOM определяет объектно-ориентированный интерфейс прикладного программирования для доступа к Web-страницам и XML-документам и их модификациям // PC-WEEK (Russian Edition).— 1997 .— № 50(124) .— С. 34–35.

19. Мусин О. Р. Трех- и четырехмерные модели для геоинформационного картографирования // Картография на рубеже тысячелетий: Докл. I Всерос. конф.