

НАУЧНЫЕ КОЛЛЕКЦИИ ИНФОРМАЦИОННЫХ РЕСУРСОВ В ЭЛЕКТРОННЫХ БИБЛИОТЕКАХ

М.Р. Коголовский
Институт проблем рынка РАН
Москва, Россия
e-mail: kogalov@cemi.rssi.ru

Абстракт

Одним из важных направлений активно развивающейся в последние годы разработки информационных систем нового класса, называемых *электронными библиотеками*, является создание для них коллекций информационных ресурсов. Коллекции представляют собой наиболее распространенную форму организации информационных ресурсов в электронных библиотеках. В связи с широкими возможностями существующих информационных технологий и разнообразием природы информационных ресурсов, характеристики различных коллекций весьма многообразны. Однако коллекции обладают и некоторыми общими свойствами, понимание которых имеет существенное значение при их разработке.

В предлагаемой работе обсуждаются основные общие свойства коллекций информационных ресурсов, методы систематизации, применяемые при их формировании, рассматриваются вопросы генезиса коллекций, роль метаданных в коллекциях, важнейшие специфические особенности научных коллекций, а также перспективные информационные технологии и стандарты, которые могут применяться для создания, поддержки и использования коллекций. Работа частично поддержана грантом РГНФ 96-02-12016.

1 Введение

Объявленная в США в конце 1993 г. по инициативе Национального научного фонда США (NSF), Агентства перспективных исследований Министерства обороны (ARPA) и Национального агентства по космическим исследованиям (NASA) программа "Digital Libraries Initiative" (DLI) стимулировала разработки в ряде стран информационных систем нового класса, ключевую роль в которых играют сформировавшиеся в 90-е годы технологии управления информационными ресурсами (ИР) и коммуникационные средства. В 1998 г. Министерством науки и технологий РФ совместно с РФФИ была объявлена аналогичная российская межведомственная программа исследований и разработок в этой области. В качестве русскоязычного эквивалента "Digital Library" стал

Первая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
19 - 21 октября 1999 г., Санкт-Петербург

использоваться термин *электронная библиотека* (ЭБ), хотя авторы DLI предпочли термин *Digital Library* варианту *Electronic Library* с тем, чтобы подчеркнуть существенное влияние цифрового представления ИР на характер технологий, применяемых в системах рассматриваемого класса [22].

Электронная библиотека рассматривается участниками программы DLI как система, обеспечивающая сообществу пользователей доступ понятным для них образом к большим репозиториям мультимедийной информации и знаний, организованным при отсутствии каких-либо сведений о способах их применения [46]. По мнению директора программы DLI С. Гриффина, ЭБ - это "средство обеспечения интеллектуального доступа к распределенным хранилищам информации на основе создания информационных сред, которые возвышают доступ от возможности получения необработанных данных - битов - до восприятия более полных знаний и смысла, содержащихся в цифровых коллекциях" [22]. В этой же работе ЭБ квалифицируются как "глобальные, многоязычные репозитории данных, знаний, звуковых образов и изображений...".

Некоторое представление о желательных функциональных возможностях ЭБ в будущем дают предложенные на организованном NSF симпозиуме в Санта Фе (1997) [45] основные направления исследований для второй фазы программы DLI (DLI-2), к числу которых были отнесены системные проблемы, проблемы управления коллекциями и проблемы, связанные с поддержкой пользователей. Концентрированно представляя эти предложения в [22], С. Гриффин указывает, что на этой стадии разработок "будет уделяться основное внимание интероперабельности и интеграции технологий, содержанию (content), разработке коллекций и управлению ими, приложениям и операционной инфраструктуре, а также пониманию ЭБ в специфических для областей применения экономических, социальных и международных контекстах, иными словами, электронным библиотекам как системам, ориентированным на человека".

На симпозиуме [45] подчеркивалось, что не следует отождествлять ЭБ с совокупностью оцифрованных коллекций и инструментария управления ими. ЭБ нужно понимать более широко, как среду, объединяющую "коллекции, сервисы и людей для поддержки полного жизненного цикла создания, распространения, использования и сохранения данных, информации и знаний".

Среди отечественных специалистов, наряду с таким пониманием ЭБ как информационной системы нового класса с указанными выше возможностями, суще-

ствуется и буквальная трактовка термина *электронная библиотека* как системы, обеспечивающей "электронизацию" традиционных библиотечных технологий с предоставлением полнотекстовых и/или мультимедийных электронных копий ИР ее пользователям. Уже сама функциональная направленность подобного рода системы, по мнению сторонников такого подхода, дает право квалифицировать ее как ЭБ, независимо от информационных технологий, на которых она базируется. Можно надеяться, однако, что развитие исследований и разработок в области ЭБ в нашей стране естественным образом приведет к единому и адекватному пониманию содержания обсуждаемого термина. Ведь аналогичная ситуация возникала в свое время и в связи с другими "импортированными" терминами в области информационных систем - *банк данных, база данных, модель данных*.

Завершая наше краткое обсуждение понятия "электронная библиотека", необходимо сделать следующие два замечания.

Во-первых, это обсуждение представляется здесь уместным по той причине, что от интерпретации смысла обсуждаемого термина существенно зависят возможности тех систем, которые он обозначает, и, в частности, требования к организации и свойствам используемых в них ИР и коллекций ИР, являющихся предметом обсуждения данной работы.

Во-вторых, следует заметить, что на формирование российской программы "Электронные библиотеки" значительное влияние оказывают представления, сложившиеся под влиянием программы DLI, которая в США имеет общенациональный статус, и в этой связи вполне естественной является ее ориентация, главным образом, на управление национальными ИР. Именно поэтому в российской программе основное внимание предполагается уделить нормативно-правовой базе отечественных ЭБ, технологиям работы с большими и сверхбольшими объемами данных, обеспечению интероперабельности ИР и глобального доступа к ним.

Между тем, ЭБ вовсе не обязательно должна быть ориентирована на глобальное сообщество пользователей, точно так же, как наряду с традиционными публичными библиотеками существуют библиотеки организаций и личные библиотеки. Эта точка зрения высказывается также, например, в [49], где аргументируется право на жизнь *персональных электронных библиотек* (Personal Digital Library). Круг пользователей конкретной ЭБ должен определяться прежде всего социальной значимостью тех ИР, которые она предоставляет, возможностями и целесообразностью их обобществления. Эти обстоятельства представляются нам весьма важными в контексте обсуждения научных ЭБ, специфического вида систем рассматриваемого класса, и используемых в них коллекций. По нашему мнению, системы ЭБ персонального и лабораторного уровня, а также уровня научного учреждения, не в меньшей мере необходимы для поддержки научных исследований, чем глобальные системы.

Создание коллекций ИР является одним из важных направлений разработки ЭБ, поскольку они представляют собой наиболее распространенную форму организации информационных ресурсов в таких системах. В связи с широкими возможностями существующих информационных технологий и разнообразием природы информационных ресурсов, характеристики различных коллекций весьма многообразны. Однако они обладают и некоторыми общими свойствами, понимание которых имеет существенное значение при их разработке.

В предлагаемой работе обсуждаются основные общие свойства коллекций ИР в электронных библиотеках, методы систематизации, применяемые при их формировании, рассматриваются вопросы генезиса коллекций, роль метаданных в коллекциях, важнейшие специфические особенности научных коллекций, а также перспективные информационные технологии и стандарты, которые могут применяться для создания, поддержки и использования коллекций.

2 Коллекции информационных ресурсов

Коллекция информационных ресурсов представляет собой систематизированную совокупность ИР, объединенных по какому-либо критерию принадлежности, например, по общности содержания, источников, назначения, по кругу пользователей, способу доступа и т.д. Коллекции являются наиболее распространенной формой организации информационных ресурсов в ЭБ.

С функциональной точки зрения, ИР коллекции подразделяются на *данные* (информацию) и *метаданные* (метайнформацию). Ресурсы первого вида представляют интересующие пользователей сведения о предметной области этой коллекции. В свою очередь, метаданные коллекции характеризуют свойства принадлежащих ей ресурсов и самой коллекции в целом как сущностей реального мира. Следует заметить, что такое разделение ИР является весьма условным. Действительно, ИР, являющиеся метаданными по отношению к некоторым другим ИР коллекции или к коллекции в целом, в некоторых случаях (когда предметная область рассматривается пользователем на метауровне) играют роль данных. Функции метаданных в коллекциях будут подробно обсуждаться далее (см. п. 5).

Систематизированный характер ИР коллекции является принципиально важным ее свойством, отличающим коллекцию от других наборов ИР. Поэтому этот вопрос заслуживает специального обсуждения (см. п. 3). Осмысленная систематизация ИР не только облегчает доступ пользователей к ним, но и дает возможность целенаправленно и рациональным образом исследовать с их помощью предметную область коллекции.

Наряду с систематизированностью ИР, к числу важных свойств любой коллекции относятся ее назначение, характеристики генезиса, способ задания состава принадлежащих ей ИР и правила их именования, виды используемых базовых информационных технологий, характеристики представления ИР (среда представления - текст, графика, и т.п., форматы, способы кодирования, используемый естественный язык и др.), однородность/неоднородность ИР (в различных аспектах), их сосредоточенность/распределенность, характеристики объема ИР, степень динамичности состава коллекции и состояния ресурсов, степень полноты коллекции, оценка непротиворечивости ИР, характеристика социальной значимости коллекции, способы доступа пользователей к ИР коллекции и др. Кратко обсудим некоторые из перечисленных свойств. Методы систематизации коллекций, вопросы их генезиса и информационных технологии, используемые для создания и использования коллекций, обсуждаются в последующих разделах.

Возможны различные подходы к *заданию состава* коллекции. В простейшем случае коллекция задается явным образом - непосредственно как совокупность принадлежащих ей ИР или как список ссылок на них (например, URL/URI ресурсов WWW). Другой подход преду-

смачивает неявное задание состава ИР путем спецификации в какой-либо форме критерия принадлежности ИР данной коллекции (Membership Criteria) [30]. Такой подход можно использовать в коллекциях, формируемых на основе глобальной распределенной гипертекстовой среды WWW. Еще одним примером его использования может служить задание коллекции, формируемой на основе полнотекстовой документальной системы путем спецификации поискового запроса. Состав ИР неявно заданных коллекций динамичен и в каждый момент времени зависит от состояния информационного пространства - источника коллекции. Помещаемые в информационное пространство или удаляемые из него ИР, удовлетворяющие критерию принадлежности коллекции, автоматически включаются или исключаются из нее. Поэтому коллекции подобного рода естественно называть *динамическими*. Конечно, для "материализации" коллекции при таком подходе необходим специальный сервис для поиска и выборки принадлежащих ей ресурсов.

На практике часто используются также коллекции, состав и состояние ИР которых неизменны или изменяются не слишком часто (*статические коллекции*). Такие коллекции обладают важным свойством - *тиражируемостью*. Примерами статических тематических коллекций гипермедийных ИР являются многие популярные Web-сайты. Широко практикуемое в WWW создание "зеркальных" сайтов имеет смысл именно в связи со статическим характером сайтов-источников.

Во многих разработках единая система именования ИР, принадлежащих коллекции, является одним из элементов ее систематизации. Выбор эффективных правил именования ИР весьма существенен, особенно для крупных коллекций. Лаконичные имена, легко ассоциируемые с обозначаемыми ими ресурсами коллекции, существенно упрощают работу пользователей. Присваиваемые ресурсам имена должны быть уникальными в рамках коллекции в целом или некоторых подмножеств ее ИР. Иногда на имя ресурса возлагается функция идентификации его (логического) места в коллекции в рамках принятой систематизации.

Природа и среда представления ИР коллекций могут быть различными. Это могут быть научные отчеты, монографии, результаты наблюдений природных феноменов, данные компьютерных модельных экспериментов или приборных измерений, художественные или музыкальные произведения, географические карты и т.п. ИР могут быть представлены в какой-либо моно-среде - тексты, изображения, аудио или видео - или являются мультимедийными. Организация и способы представления ИР коллекции, в том числе, и метаанных, существенным образом зависят от информационных технологий, используемых в данной ЭБ (см. п. 7).

Важными свойствами коллекций являются также *полнота* и *непротиворечивость* содержащихся в ней ресурсов. Полнота ИР является относительным свойством. Оценивать полноту ИР коллекции можно лишь в контексте ее назначения и положенных в ее основу принципов систематизации. Что касается непротиворечивости ИР, то методы, необходимые для ее обеспечения, существенным образом зависят от характера самих ИР. Например, для поддержки непротиворечивости фактографических ИР могут быть использованы автоматизированные механизмы, традиционно применяемые в системах баз данных. В большинстве коллекций слабоструктурированных данных, основанных на Web-технологиях, не обеспечивается какой-либо автоматизации, предназначенной для этих

целей.

3 Методы систематизации коллекций

Систематизация коллекции осуществляется, как правило, на основе свойств предметной области (ПО) коллекции и/или свойств составляющих ее ИР и их пользователей, характеристик процесса формирования коллекции или участников этого процесса, по хронологическому принципу и т.д.

Систематизация коллекции с использованием свойств ее ПО должна основываться на *концептуализации* ПО. Построенная в результате изучения и анализа ПО концептуальная модель может вместе с тем использоваться для обеспечения и оценки полноты содержания коллекции.

Концептуальная модель ПО может иметь различные формы представления. При реализации коллекции на основе технологии баз данных эта модель представляется в форме *концептуальной схемы*, которая далее отображается в среду конкретной выбранной инструментальной СУБД и описывается схемой базы данных. Такая модель описывает типы сущностей, представляющие интерес (их атрибуты и, возможно, поведение), связи между типами сущностей и ограничения целостности, которым должны удовлетворять экземпляры типов и связей.

Для многих коллекций концептуальная модель ПО представляется в форме *классификатора*, определяющего одномерное или многомерное пространство классификационных признаков. В таком случае отдельные (не обязательно все) точки или гиперплоскости этого пространства представляют классы сущностей ПО, соответствующие ресурсам коллекции. В качестве классификационных признаков обычно используются наиболее существенные свойства сущностей ПО коллекции. Как правило, для классификационных признаков может быть предопределено множество их возможных значений. При этом множество значений отдельного признака может иметь линейную или иерархическую структуру.

Многомерные линейные и иерархические классификаторы широко применяются, в частности, для систематизации коллекций временных рядов макроэкономических показателей (где используются, например, классификаторы отраслей хозяйства, форм собственности, экономических районов, выпускаемой продукции и т.п.), различного рода коллекций в систематизирующих областях науки (ботанике, зоологии и др.) и во многих других случаях.

Одномерные линейные и иерархические классификаторы весьма часто используются также в коллекциях гипермедийных и гипертекстовых ресурсов, основанных на Web-технологиях.

Для удобства работы с коллекцией и ее статистической обработки при использовании классификаторов часто применяется кодирование значений классификационных признаков с помощью порядковых, диапазонных или позиционных кодов. В случае позиционного кодирования значения кодов позволяют отражать иерархические отношения между значениями признаков, обеспечивая простоту процедур агрегирования данных в статистических коллекциях.

Заметим, что с некоторыми коллекциями, например, научными коллекциями экспериментальных данных, может ассоциироваться несколько сосуществующих концептуальных моделей ПО, соответствующих различным теориям исследуемого феномена. В таких случаях ис-

пользование разных моделей приводит к различным интерпретациям одних и тех же ИР коллекции.

Детальное рассмотрение других подходов к систематизации коллекций не входит в задачу данной работы.

4 Генезис коллекций

В процессе формирования коллекции необходимо решить ряд задач. Основные из них - определение содержательного состава коллекции, источников ее формирования, принципов систематизации коллекции, обеспечение полноты и непротиворечивости ИР, выбор подходящих информационных технологий для формирования, поддержки и использования коллекции.

Задачи, связанные с определением состава коллекции, ее систематизацией и обеспечением полноты ИР, взаимосвязаны. Основой их решения является чаще всего концептуализация предметной области коллекции, о которой уже шла речь в п. 3.

Для формирования коллекций ИР могут использоваться разнообразные источники: периодические издания; монографии; издания художественной литературы; научные отчеты; диссертации; музейные экспонаты; натурные наблюдения и измерения характеристик процессов и явлений в природе, в социально-экономической среде или в технических системах, а также данные, полученные в результате обработки таких измерений; результаты компьютерного моделирования; другие уже существующие коллекции и т.д.

В некоторых случаях включаемые в формируемую коллекцию ИР уже существуют в цифровом виде как автономные ресурсы или в составе других коллекций - систем баз данных, документальных систем, различного рода мультимедийных информационных систем, Web-сайтов или информационных систем, основанных на интеграции каких-либо из перечисленных технологий.

Если электронные источники ИР коллекции существуют, то имеется два подхода. При первом из них новая коллекция может создаваться как собрание ресурсов, существующее автономно от их источников. В таких случаях, как правило, возникают проблемы поиска и выбора требуемых ИР из коллекций-источников, трансформации их представлений, в соответствии с требованиями технологической среды формируемой коллекции.

При втором подходе новая коллекция создается как "виртуальная", без порождения дополнительных копий заимствуемых ИР. Такая коллекция может быть задана неявным образом (см. п. 2) либо как список ссылок на составляющие ее ИР в среде-источнике. Случай неявного задания обладает рядом достоинств, но могут возникнуть достаточно сложные проблемы интеграции неоднородных данных, которые в последние годы активно исследуются многими коллективами, например, согласование и обеспечение непротиворечивости ИР, заимствуемых из многих источников. Подобные проблемы хорошо известны в области неоднородных распределенных баз данных и хранилищ данных (data warehouse), где для их решения используют конверторы данных, медиаторы, адапторы а также различные методы "очистки" данных (data scrubbing, data cleansing).

Если коллекция формируется в течение продолжительного времени и/или включает ИР, относящиеся к различным временным периодам существования ПО, то в силу происходящих в ПО изменений или появления новых знаний о ней у создателя коллекции могут потребоваться существенные изменения набора значений класси-

фикационных признаков тех сущностей ПО, сведения о которых накапливаются в коллекции. Более того, может возникнуть необходимость в радикальном изменении положенной в основу коллекции классификационной схемы.

Указанные проблемы исследовались в технологии баз данных. Они связаны с темпоральными свойствами данных во временных базах данных, а также с эволюцией схемы базы данных. Серьезные трудности связаны при этом, в частности, с необходимостью поддержки индивидуальности сущностей ПО, изменяющихся во времени (естественный подход к решению этой проблемы возможен в объектных базах данных), с обеспечением сопоставимости значений свойств таких сущностей в разные периоды времени. Проблемы такого рода возникают, например, в базах данных временных рядов макроэкономических показателей при осуществлении структурных сдвигов в экономике и изменениях хозяйственного механизма, при формировании коллекций экспериментальных данных в различных областях науки, где также необходимо обеспечить сопоставимость результатов экспериментов, проводившихся в различное время и при изменяющихся условиях.

Наряду с ситуациями, когда включаемые в коллекцию ИР уже существуют в цифровом представлении, возможна и ситуация, когда необходима их предварительная оцифровка. Как правило, для этих целей могут использоваться типовые широко распространенные технологии. Однако иногда требуется специальное оборудование и программное обеспечение для этих целей. Так, при автоматической регистрации показаний различных научных приборов необходимы специальные интерфейсные средства для сопряжения таких приборов и средств вычислительной техники с использованием преобразователей аналоговых сигналов в цифровые, а также специализированное программное обеспечение.

Информационные технологии, используемые для создания и поддержки коллекций ИР, подробно рассматриваются в п. 7.

5 Метаданные в коллекциях ИР

Конкретные функции метаданных коллекций и их содержание могут значительно различаться в зависимости от характера ЭБ и конкретной коллекции. Однако в любом случае метаданные должны определять состав коллекции, обеспечивать корректную интерпретацию и обработку представленных в ней информационных ресурсов механизмами ЭБ. В связи с этим они должны описывать критерии принадлежности ИР данной коллекции (для неявно заданных коллекций, см. п. 2), структурные (а при использовании объектной парадигмы моделирования предметной области и поведенческие) свойства ИР (типы, связи), форматы их представления, место размещения, ограничения целостности, ограничения управления доступом и т.п.

Вместе с тем, метаданные могут описывать и семантические свойства ИР, например, содержание ресурсов, сведения об их генезисе, авторах, контекстах применения, о классификационной системе коллекции и схеме именования ее ИР, о единицах измерения значений.

Состав метаданных этой группы существенно зависит от генезиса и природы ИР коллекции. Действительно, для ресурсов экспериментальных данных метаданные должны описывать вид эксперимента, характеристики используемых приборов, условия и время проведения

эксперимента, а также, возможно, математическую модель исследуемого феномена. Для коллекций по собраниям музейных экспонатов необходимы совершенно иные метаданные, например, виды экспонатов, сведения об их авторах, местонахождении, времени и месте происхождения, о проведенных реставрационных работах и т.д.

С метаданными коллекций связана достаточно сложная проблема, уже упоминавшаяся в п. 4 и заключающаяся в следующем. В период существования коллекции в ее предметной области могут происходить такие изменения, которые приводят к необходимости соответствующих изменений в систематизации коллекции, например, изменений в используемом классификаторе ИР. Такие ситуации характерны для систем баз данных с эволюционирующей схемой.

Типичным примером коллекций, для которых такие явления являются правилом, а не исключением, могут служить коллекции временных рядов экономических показателей. В них по необходимости должны находить отражение структурные сдвиги в экономике, изменения хозяйственного механизма, реорганизации экономических объектов, изменения номенклатуры выпускаемой продукции. В связи с этим приходится изменять сами системы измеряемых экономических показателей, подвергаются изменениям классификаторы, лежащие в их основе (а, следовательно, и в основе рассматриваемых коллекций), возникают сложности, связанные с несопоставимостью старых и новых значений показателей. В таких условиях необходимо обеспечить корректное одновременное и, возможно, совместное использование данных, относящихся к старой и новой структуре классификатора.

В ЭБ могут использоваться различные способы и средства представления метаданных коллекций, в зависимости от применяемых информационных технологий. Эти технологии базируются на целом ряде международных, национальных и промышленных стандартов. Основные из них рассматриваются ниже в п. 7.1.

6 Многообразие и особенности научных коллекций

Научные коллекции ИР весьма разнообразны. Наряду с общими свойствами, присущими любым коллекциям, они обладают во многих случаях и специфическими особенностями. Эти разнообразие и специфика являются следствиями не только многообразия сфер исследований и особенностей исследовательской деятельности, но и различий в методологии и "технологиях" исследований в разных областях науки. Отсюда, в свою очередь, возникает потребность в широком спектре информационных технологий для создания научных ЭБ и коллекций.

Научные коллекции различаются, прежде всего, масштабом круга пользователей. Имеют право на жизнь как персональные коллекции отдельных ученых и коллекции коллективов исследовательских лабораторий, так и коллекции, имеющие национальную или мировую значимость и предоставляемые для глобального доступа. Важно заметить, что эта характеристика коллекции может изменяться с течением времени. Персональная коллекция может со временем приобрести высокий уровень значимости для многих ученых, и в этой связи потребуются обеспечить к ней более широкий доступ.

По сравнению с крупномасштабными коллекциями, уже в силу этого значительно более консервативными, персональные и лабораторные научные коллекции могут

иметь относительно короткий срок жизни, обычно они более динамичны по структуре и составу ИР.

Информационные потребности исследователя вообще значительно более динамичны по сравнению, например, с относительно стабильными потребностями управленческих работников, чья деятельность в значительной мере регламентирована. Поэтому в научных ЭБ должны предусматриваться весьма гибкие пользовательские интерфейсы, способные адаптироваться к изменению потребностей пользователей. Это обстоятельство должно учитываться при организации научных коллекций.

В отличие от ИР других видов коллекций, в научных коллекциях могут быть представлены сведения не о реальных процессах и явлениях, а гипотетические данные или данные компьютерных экспериментов с гипотетическими исследовательскими моделями.

Достоверность сведений, содержащихся в ИР коллекции, может быть обеспечена лишь относительно достигнутого уровня знаний в рассматриваемой области науки. Она может быть опровергнута в процессе дальнейших исследований.

В научных коллекциях могут содержаться альтернативные (возможно, даже противоречивые) сведения об исследуемых объектах, явлениях или процессах. В таких ситуациях, конечно, неправомерно вести речь об интегральной целостности данных в коллекции.

При разработке коллекций ИР в естественно-научных исследованиях (например, в механике, теоретической физике) иногда применяется подход, когда для одной и той же коллекции сосуществует несколько концептуализаций предметной области, в соответствии с различными теориями и моделями исследуемого феномена. Это позволяет, в частности, различным образом интерпретировать данные одних и тех же экспериментов.

Для научных коллекций ИР не является необычной неполнота и/или нечеткость представляемых ими сведений. Более того, сама концептуализация предметной области исследования, положенная в основу коллекции, может иметь гипотетический характер.

Особенности коллекций в конкретных областях науки проявляются в преобладании некоторых видов ИР, в нетрадиционном характере их обработки, в предъявляемых к коллекциям каких-либо специфических требованиях. Проиллюстрируем это некоторыми примерами.

Коллекции в систематизирующих научных дисциплинах (ботаника, зоология, минералогия и др.) обычно основываются на классификаторах различного рода, сама разработка которых является существенным элементом проводимого научного исследования. Динамика таких коллекций, как правило, является однонаправленной - они лишь пополняются. Эти коллекции могут тиражироваться (например, на компакт-дисках), иногда вместе с программными средствами доступа к ним. Для коллекций такого рода важное значение имеют возможности поиска и получения из них выборок по различным критериям, последовательного просмотра их элементов, относящихся к заданному разделу классификатора, а также статистического анализа их состава.

В областях науки, где отводится важное место компьютерному моделированию (некоторые разделы математики, механики, физики, биологии, экономико-математические исследования и др.), организация коллекций должна быть приспособлена для использования их ИР в различных модельных компьютерных экспериментах. Такие коллекции обычно не только служат источниками исходных данных для экспериментов, но и

сохраняют характеристики самих проведенных экспериментов и их результаты. Здесь важно также обеспечить средства предоставления пользователю данных из коллекций в форме, удобной для интерпретации исходных данных и результатов исследований.

Коллекции ИР в некоторых областях наук о Земле (в геофизике, океанологии, физике атмосферы), в экономических исследованиях должны поддерживать пространственные и/или темпоральные свойства данных. Составными частями таких коллекций часто являются временные и/или пространственные ряды наблюдений. При разработке таких коллекций и механизмов ЭБ, обеспечивающих доступ к ним, целесообразно базироваться на пространственно-временных моделях данных.

В космических исследованиях коллекции характеризуются, как правило, огромными объемами данных, значительную часть которых составляют оцифрованные данные радиотелеметрии.

В таких областях науки, как география, экология, демография, региональная экономика, значительную часть ИР коллекций составляют картографические данные. Работа с такими коллекциями требует использования технологий ГИС.

Основными видами ИР в химических коллекциях являются графически представленные структурные формулы соединений, спектрограммы, текстовые описания качественных свойств веществ и их количественные характеристики.

Разнообразный характер имеют ИР исторических коллекций. Так, исторические коллекции Национальной электронной библиотеки США [3], создаваемой на основе фондов Библиотеки конгресса, включают оцифрованные фотографии, факсимиле редких книг, звукозаписи, карты, кинодокументы и видеозаписи, разнообразные текстовые документы и т.п.

Отметим, наконец, что во всех областях научных исследований применяются библиографические коллекции, а также коллекции полнотекстовых научных публикаций. Значительный эффект может дать интеграция коллекций этих двух видов. В качестве примера можно привести общедоступную в среде WWW библиографию М. Лея из Трирского университета по системам баз данных и логическому программированию [32], охватывающую публикации в многочисленных журналах и трудах конференций, а также монографии по этой проблематике. По результатам поиска в библиографии пользователь может, не прерывая сеанса работы с поисковым сервисом, получить доступ к аннотациям и/или полным текстам найденных работ, содержащимся в электронных библиотеках АСМ и IEEE, а также ряда издательств, если он обладает необходимыми полномочиями.

7 Информационные технологии создания, поддержки и применения научных коллекций

В разработке научных коллекций ИР для ЭБ наряду с традиционными технологиями баз данных, документальных систем, а также Web-технологиями, находят применение новые подходы, формирующиеся в каждом из этих направлений, предпринимаются попытки создания смешанных интегрированных технологий. Начинают находить применение некоторые важные ранее разработанные технологии, например, технологии интероперабельности CORBA [11], которые уже достаточно широко применяются в разработках крупных корпоративных информационных систем и в других областях. Важное значение

в разработке коллекций имеют также технологии управления метаданными.

В этой работе не преследуется цель детального и всестороннего анализа состояния развития информационных технологий, которые могут быть использованы для создания, поддержки и применения коллекций ИР для ЭБ. Мы ограничимся здесь кратким обсуждением наиболее важных из них, прежде всего, технологий промышленного характера либо близких к достижению такого статуса, стандартов, на которых они основаны, а также некоторых новых областей их применения для создания научных коллекций.

7.1 Технологии управления метаданными

Как уже отмечалось выше, в настоящее время разработаны и используются разнообразные подходы к управлению метаданными информационных систем, создан целый ряд международных, национальных и промышленных стандартов в этой области, многие из которых могут найти применение в коллекциях ИР электронных библиотек. Ниже мы кратко рассмотрим наиболее перспективные из них.

При использовании инструментария баз данных структурные метаданные и метаданные, описывающие ограничения целостности данных, представляются в схеме базы данных, специфицируемой средствами языка описания данных СУБД.

Для более интенсивной поддержки метаданных в системах баз данных и других типах информационных систем, разрабатываемых с помощью инструментальных средств анализа и проектирования, существует разработанный и развиваемый ISO/IEC JTC1 международный стандарт *Information Resource Dictionary System (IRDS)* [26]. Он описывает системы, которые предназначены для создания и поддержки справочника ИР организации, для обеспечения доступа к нему, а также средства определения представленных в этом справочнике ресурсов. Такой справочник может содержать, например, информацию об используемых организацией данных, о процессах, связанных с управлением этими данными, о необходимом для этого оборудовании, о лицах, ответственных за поддержку такой информации. Предусматривается многоуровневая архитектура моделирования метаданных в справочнике с отображением их в конечном счете в базу данных. В стандарте специфицирована среда систем справочников ИР, интерфейсы предоставляемых ими сервисов, связывания для языков программирования C и Ada. Проводятся работы по интеграции этого стандарта с технологией OMG CORBA [11], созданы спецификации средств экспорта/импорта для IRDS, обеспечивается поддержка возможностей именованного ресурсов и тезаурусов. Стандарт IRDS используется для создания репозитория метаданных в системах баз данных и хранилищах данных. По отношению к коллекциям ИР в ЭБ он может рассматриваться как средство низкого уровня.

В распределенных неоднородных интероперабельных объектных средах, основанных на архитектуре CORBA, для поддержки метаданных в 1997 г. OMG был принят стандарт *Meta Object Facility (MOF)* [38]. Этот стандарт основан на объектной модели, базирующейся на концепциях известной модели "сущностей-связей" П. Чена [9] и являющейся расширением модели-ядра OMG, на которую опирается стандарт CORBA. В стандарте MOF предложены спецификации отображения модели MOF в язык определения интерфейсов IDL стандарта CORBA

[11, 64], а также спецификации основанных на CORBA сервисов для управления метаинформацией. Предполагается, что модель MOF будет играть роль мета-мета модели, служащей для описания мета-моделей, которые лежат в основе различных средств объектного анализа и проектирования. Разработчиками MOF предложены также две альтернативные нотации для модели MOF - графическая нотация языка UML (см. ниже) и язык MODL (Meta-Object Definition Language), описанный в спецификациях MOF, однако, формально не являющийся частью этого стандарта. В настоящее время OMG MOF представлен в ISO для придания ему статуса официального международного стандарта.

Важным средством для представления метаданных в ЭБ может стать язык *UML (Unified Modeling Language)* [53, 59]. Этот язык был принят OMG одновременно с MOF в сентябре 1997 г. в качестве индустриального стандарта, призванного обеспечивать interoperability объектно-ориентированных инструментальных средств анализа и проектирования систем, опирающихся на архитектуру CORBA.

UML создан известными специалистами в области объектного анализа и проектирования Г.Буча, И.Якобсона и Д.Рэмбо (компания Rational Software). Он синтезирует и развивает подходы, реализованные в ранее разработанных ими методах (метод Буча, OMT, OOSE), учитывает возможности других получивших широкое признание методологий, является независимым от конкретных используемых при реализации проектируемых систем языков программирования и может быть адаптирован к различным технологическим процессам разработки.

Спецификации стандарта UML включают описание семантики языка, его графической нотации, а также расширений языка для процесса разработки программного обеспечения Objectory (предложенного И.Якобсоном) и для моделирования деловых приложений. В UML предусматривается возможность описания ограничений, налагаемых на объекты и ассоциируемых с графическими моделями. Для декларации таких ограничений в стандарте вводится объектный язык ограничений - OCL (Object Constraint Language).

Сообщество разработчиков технологий для WWW также серьезно озабочено необходимостью явного введения метаданных в этой среде с тем, чтобы повысить эффективность поиска ресурсов поисковыми машинами Web, обеспечить основу для решения проблем семантической интеграции распределенных информационных ресурсов и их повторного использования.

Ранние предпринятые в этом направлении попытки привели к включению в версию языка *HTML 2.0* простейших средств, позволяющих встраивать метаданные в HTML-документы [1]. Предполагалось, что содержание документа будет характеризоваться значениями некоторых атрибутов, вообще говоря, различных для разных документов. Описания семантики наборов таких атрибутов для различных предметных областей, называемые схемами, должны быть представлены на каких-либо WWW-серверах. Входящие в набор атрибуты называются элементами соответствующей схемы. В синтаксис языка был введен новый тег *META* с двумя атрибутами - *NAME* и *CONTENT*. Первый из них задает имя элемента схемы, квалифицированное идентификатором схемы, а второй - его значение. Теги *META* могут повторяться произвольное число раз в HTML-документе, позволяя тем самым ассоциировать с ним необходимое количество

атрибутов метаданных. Ссылка на местоположение схемы в WWW (URL) вместе с присвоенным ей идентификатором указывается в теге *LINK*. Таким образом, появилась возможность включать в HTML-документы структурированные метаданные, характеризующие их содержание, например, значения элементов Дублинского ядра [58].

Впоследствии указанные средства языка HTML получили дальнейшее развитие в его версии *HTML 4.0* [24] под влиянием подготовленного W3C (World Wide Web Consortium) к этому времени проекта Resource Definition Framework (RDF) средств описания семантики документов в среде Web, основанных на новом языке разметки XML. В частности, для тега *META* были введены дополнительные атрибуты *LANG* и *SCHEMA*, которые позволяют задать, соответственно, язык представления значения элемента метаданных в этом теге и уточняющий контекст для адекватной его интерпретации. Появился также новый атрибут профиля документа *PROFILE* в теге *HEAD* заголовка, содержащий ссылку (URL) на ресурс в Web, в котором определяются элементы метаданных данного документа и их значения. Формат содержания профиля в языке не регламентирован.

Проблема разработки средств для спецификации метаданных возникла и в связи с созданием новых Web-технологий, основанных на языке разметки XML [19] - стандарте консорциума W3C. Прежде всего, некоторые возможности для этой цели были предусмотрены в самом языке XML. Их называют декларацией типа документов. Спецификация этих деклараций средствами XML называется *Document Type Definition (DTD)* и позволяет описывать допустимые структуры гипертекстовых документов рассматриваемой категории в терминах составляющих их элементов.

Для каждого типа элементов документа указывается вид их содержания (содержания нет, литерная строка, список вложенных элементов, смешанное содержание - литерная строка и вложенные элементы), обязательно или факультативно присутствуют элементы в документе, а также перечень соответствующих им атрибутов и их типы. Соотношение между DTD и множеством соответствующих XML-документов аналогично соотношению между схемой базы данных и множеством описываемых ею конкретных баз данных. Верификация конкретного документа на соответствие спецификации DTD может осуществляться процессорами языка XML, например, поддерживающим этот язык Web-браузером.

DTD может использоваться двумя способами. В простейшем случае эти спецификации встраиваются непосредственно в XML-документ. Спецификации DTD для категории документов, представляющей интерес для многих пользователей, могут размещаться на каком-либо Web-сервере для общего доступа, а в конкретных XML-документах делаются ссылки на него. Средства DTD уже находят применение для создания научных коллекций. Так, разработаны DTD для исторических коллекций в Библиотеке конгресса США [2], для описания астрономических инструментов [12] и др.

В настоящее время W3C проводит работы по дальнейшему развитию средств описания структуры и других свойств XML-документов. С этой целью создается стандарт языка определения схемы для XML-документов - *XML Schema Definition Language (XML Schema)*, проект которого был недавно опубликован [55, 56].

Наиболее важные новые возможности XML Schema, по сравнению с DTD, заключаются во введении более

развитой совокупности типов значений атрибутов элементов XML-документов, в допущении наряду с закрытой моделью спецификаций DTD также и открытой модели, при которой пользователь может дополнять повторно используемую схему новыми спецификациями и т.д.

Поскольку язык XML Schema является приложением XML, специфицированная его средствами схема сама является XML-документом. Схема таких схем может быть использована для верификации конкретных схем. Одна из версий спецификации схемы схем приведена в проекте стандарта как его составная часть.

Под влиянием и на основе исследований, проводимых в рамках программы DLI, консорциум W3C принял также стандарт средств для описания семантики IP в среде Web, независимых от конкретной предметной области, - *Resource Definition Framework (RDF)*. Этот стандарт состоит из двух частей. В первой из них [47] предлагается семантическая модель и синтаксис базирующегося на XML языка, средствами которого может быть выражена семантика IP - RDF-спецификация. Использование XML как языковой среды для представления RDF-спецификаций естественным образом решает проблему обмена метаданными в WWW и их повторного использования в приложениях, основанных на XML.

Описание семантики IP в терминах модели RDF по существу эквивалентно ER-диаграмме [9] и декларирует множество IP, с каждым из которых ассоциируются пары "свойство - значение". Значения свойств задаются литерально либо ссылками на другие ресурсы, которые представляются, в свою очередь, их свойствами. Таким образом, свойства могут определять и связи между ресурсами. IP идентифицируются уникальным образом с помощью их URI (Uniform Resource Identifier, обобщение концепции URL в WWW). Они могут также представлять собой коллекции других IP или литералов, называемые контейнерами. Допускаются контейнеры типа мультимножества, последовательности и альтернативы.

Для того, чтобы RDF-спецификация семантики IP была полной, необходимо ассоциировать с ней описание семантики используемых в спецификации свойств, называемое в терминологии RDF *схемой*. Никаких ограничений на способ представления схемы не налагается. Достаточно лишь представить ее как некоторый ресурс в WWW, и использовать URI этого ресурса для ссылки на нее в RDF-спецификации. Характер спецификаций, глубина описания семантики свойств в схеме и степень ее формализованности, должны соответствовать потребностям приложений XML, оперирующих конкретной категорией IP, которые описываются данной RDF-спецификацией и этой схемой.

В стандарте RDF предусматривается два способа задания схем. Первый из них, более простой, состоит в использовании в качестве схемы пространства имен свойств XML - *XML-Namespaces*. Спецификации пространства имен [41] - это еще один принятый W3C стандарт в инфраструктуре информационной среды, основанной на языке XML.

Пространство имен определяет некоторый набор слов, используемых в качестве имен в XML-спецификации, и описывает семантику каждого из них. Поскольку идентификация пространств имен с помощью URI уникальна в WWW, определенные в них имена при квалификации их идентификатором пространства имен (такой идентификатор ассоциируется со ссылкой на пространство имен в XML-документе, в частности, в RDF-спецификации)

также являются глобально уникальными в WWW. Благодаря этому возможно в одной RDF-спецификации использовать имена свойств, которые принадлежат различным пространствам имен и тем самым имеют различный смысл, не опасаясь коллизий между ними.

Другой, семантически более богатый способ задания схемы, предлагаемый W3C, предусматривает использование средств RDF Schema [48] - второй части стандарта RDF. Процесс ее рассмотрения и принятия находится в настоящее время в завершающей стадии. RDF Schema предоставляет средства не только для моделирования и описания семантики свойств IP, но и для спецификации ограничений целостности.

Схема в RDF Schema представляет собой описание специфической для конкретной предметной области совокупности ресурсов RDF, которые используются для описания свойств других ее ресурсов. В связи с этим RDF-схема представляется как RDF-спецификация средствами синтаксиса, предложенного в [47]. В этой спецификации используется специальное предопределенное пространство имен.

Спецификации RDF Schema основаны на модели, близкой по ее возможностям к моделям представления знаний и использующей объектную парадигму. Эта модель опирается на концепции классов, свойств и ограничений, ассоциируемых с классами и свойствами. Для классов поддерживается иерархическое отношение класс-подкласс. Используемая в RDF Schema модель является фактически метамоделью по отношению к модели, лежащей в основе RDF-спецификаций. Средствами этой модели в стандарте [48] определяется *схема-ядро*, в терминах которой описываются конкретные схемы. Схема-ядро неявно включается в состав каждой из них. Ее ресурсами является небольшой набор "строенных" классов, свойств и ограничений целостности.

Метаданные, представленные средствами RDF, могут использоваться для более эффективного поиска ресурсов поисковыми машинами WWW, в электронных библиотеках, в описаниях коллекций страниц Web, составляющих некоторый виртуальный документ, для представления содержания IP в конкретных предметных областях, а также для поддержки различных других Web-приложений, нуждающихся в семантической информации о ресурсах.

Как уже отмечалось, в задачу RDF не входит стандартизация каких-либо наборов семантических свойств, и они могут быть различными для различных приложений. Для некоторых предметных областей уже существуют такого рода стандарты. Так, например, для описания семантики электронных текстовых документов в среде Web предложен набор свойств, названный Дублинским ядром [57] (см. ниже). В стандарте RDF показано, каким образом Дублинское ядро может быть выражено средствами RDF Schema [48].

Работы по стандартизации набора семантических свойств с ориентацией, главным образом, на публикуемые в WWW текстовые документы, заметно активизировались после основополагающего симпозиума, организованного в Дублине (США, штат Огайо) силами Online Computer Library Center и National Center for Supercomputing Applications (1995). Целью симпозиума было обсуждение состава элементов метаданных, которые могли бы использоваться для описания содержания IP, представленных в WWW, и тем самым обеспечивали бы более эффективный поиск требуемых ресурсов, а также поддержку других Web-приложений. Выработан-

ный на симпозиуме подход стал называться *Дублинской инициативой* [57].

Предложенное первоначальное множество из 13 элементов метаданных получило название *Дублинского ядра* (Dublin Core, DC). Его развитие поддерживается специально созданными органами - Директоратом Дублинского ядра, Консультативным комитетом по политике и Техническим консультативным комитетом. Конкретная работа по выработке предложений, связанных с развитием спецификаций, ведется рядом рабочих групп. Указанные организации взаимодействуют с Internet Engineering Task Force и National Information Standard Organization с целью придания DC статуса стандартов, принятых этими учреждениями.

Текущая версия спецификаций Дублинского ядра - DC 1.1 [16] включает 15 элементов. К их числу относятся: Title (Название ресурса), Creator (Лицо, организация или служба, ответственная за подготовку содержания ресурса), Subject (Тема, обсуждаемая в содержании ресурса), Description (Описание содержания ресурса в свободной форме), Publisher (Лицо, организация или служба, обеспечивающая доступ к ресурсу), Contributor (Другие участники подготовки содержания ресурса помимо указанного в Creator), Date (Дата создания или предоставления доступа к ресурсу), Type (Жанр, категория или другие характеристики природы ресурса), Format (Характер представления ресурса), Identifier (Точная ссылка на ресурс), Source (Ссылка на источник, из которого процирован данный ресурс), Language (Язык представления ресурса), Relation (Ссылка на ресурс, связанный с данным), Coverage (Область пространства, времени и т.д., к которой относится содержание ресурса), Rights (Права интеллектуальной собственности на ресурс и т.п.). Напомним, что для элементов Дублинского ядра средствами стандарта RDF [48] может быть специфицирована схема с целью использования ее в контексте RDF-спецификаций для соответствующего класса XML-документов в WWW.

В настоящее время обсуждаются направления дальнейшего развития DC и содержание следующей его версии DC 2.0 [58]. Предполагается, в частности, пересмотреть состав элементов DC, расширить возможности спецификации семантики документов в различных предметных областях путем введения уточнений (квалификаторов) для самих элементов DC и их значений, стандартизовать семантику и методы уточнений. Обсуждается, например, целесообразность замены тройки элементов Creator, Contributor и Publisher одним более общим элементом Agent, а три существующих указанных элемента могут выражаться как его подтипы. Аналогично, элемент Source может быть выражен с помощью уточнения элемента Relation. Элемент DATE также может иметь разные подтипы, позволяющие отражать даты различных событий в жизненном цикле ИР. Для некоторых элементов DC должны допускаться составные значения.

Наряду с созданием средств представления метаданных, которые описывают семантику ИР, входящих в коллекции, большое значение имеют разработки стандартов спецификаций обмена метаданными между различными инструментальными средствами разработки. Необходимость в таком обмене возникает в связи с попытками повторного использования ИР коллекций, а также при решении задач реинженерии использующих их систем.

Одна из ранних попыток в этом направлении привела еще в 1987 г. к принятию разработанного силами Electronic Industries Association (EIA) стандарта *CASE*

Data Interchange Format (CDIF) [7] обмена метаданными между инструментальными средствами CASE. В этом проекте предполагалось использовать специально созданную мета-модель, основанную на парадигме моделирования "сущностей-связей" [9]. В дальнейшем работу по развитию CDIF продолжила Рабочая группа ISO/IEC JTC1/SC7/WG11. В настоящее время CDIF представляет собой семейство стандартов ISO, независимых от поставщиков инструментальных средств CASE и от используемого в них метода проектирования. Большая часть стандартов семейства уже принята, разработка остальных завершается.

Другой группой разработчиков стандарта аналогичного назначения стал образованный в 1995 г. консорциум *Meta Data Coalition (MDC)*, в состав которого входят пользователи и такие крупные поставщики программного обеспечения, как IBM, NCR, Sybase, SAS, Platinum и др. Первая версия стандарта *Meta Data Interchange Specification (MDIS)* была принята в 1996 г. [39], и он продолжал развиваться. В процессе этой работы были установлены связи с ISO, W3C. В состав MDC вошла компания Microsoft. В апреле 1999 г. было объявлено об объединении усилий MDC и OMG для развития стандартов метаданных. Стандарт MDIS основывается на спецификации мета-модели *Open Information Model (OIS)* [43], разработанной участниками коалиции.

Нужно, наконец, упомянуть разработки в области стандартизации обменного формата для метаданных в среде XML, проводимые OMG. Их результатом стало принятие в марте 1999 г. спецификаций языка *XML Metadata Interchange (XMI)* [54]. Назначение XMI состоит в том, чтобы обеспечить простой обмен метаданными между инструментальными средствами моделирования, поддерживающими язык UML [53, 59], и репозиториями метаданных, основанными на стандарте OMG MOF, в распределенных неоднородных средах, соответствующих стандарту CORBA. Стандарт XMI предусматривает возможности обмена метаданными в режиме потока и в форме файлов стандартного формата, специфицированных на языке XML.

Подводя итоги данного раздела, следует констатировать, что в настоящее время проводятся работы по стандартизации управления метаданными, в частности, содержания поддерживаемых метаданных, методов их представления, а также обмена метаданными между различными инструментальными средствами. Применяемые для этого подходы различаются поддерживаемым ими уровнем абстракции (мета уровень, мета-мета уровень и т.п.) и функциональной направленностью метаданных, специфицируемых их средствами. Некоторые из подходов имеют специфические области применения, остальные являются, по существу, альтернативными друг другу. Выбор того или иного подхода - задача разработчика коллекции ИР, и решение ее должно быть согласовано с характером технологий, используемых для создания и применения данной коллекции в ЭБ.

7.2 Технологии баз данных

В области технологий баз данных наиболее существенное влияние на работы по созданию научных коллекций, по нашему мнению, могут оказать четыре важных фактора - (1) развитие объектной инфраструктуры, прежде всего, благодаря деятельности по стандартизации, проводимой OMG и ODMG, активизация производства "чисто"объектных СУБД; (2) появление объектно-

реляционных серверов баз данных, созданных ведущими поставщиками инструментальных средств для систем баз данных; (3) создание новых Web-технологий и интеграция технологий баз данных с Web-технологиями; (4) стандартизация интерфейсов между системами баз данных и системами программирования на языке Java.

Хорошо известны те трудности, которые возникают при попытках использования доминирующих в области баз данных реляционных технологий в ряде нетрадиционных областей применения, в том числе, и в создании научных коллекций. Эти трудности связаны, главным образом, с примитивностью поддерживаемых реляционными СУБД типов данных и отсутствием средств расширения системы типов, с невозможностью в этой связи эффективно поддерживать сложные структуры данных. В реляционных системах также нет возможности поддерживать темпоральные свойства данных, их многоверсионность.

По этим причинам для создания и поддержки коллекций в некоторых областях применения приходится создавать специальные модели данных, адекватные потребностям разрабатываемых приложений, и реализовать их как более высокий уровень представления данных над реляционной моделью. Такие модели в некоторых случаях строятся как расширения реляционной модели. Этот подход был использован автором при разработке инструментария для создания и поддержки коллекций временных рядов экономических показателей. Разработанная модель данных временных рядов конструировалась как расширение традиционной реляционной модели и была реализована в ее среде [28, 65].

Ряд указанных трудностей снимается благодаря использованию *объектных СУБД*. Такие системы позволяют вводить новые типы данных, определяемые пользователем, допускают сложное структурирование данных, хорошо интегрируются в существующую развитую инфраструктуру распределенных объектных сред, обеспечивают высокую производительность труда разработчиков приложений за счет использования API для мощных объектных систем программирования и повторного использования программного кода, а также применения существующих развитых средств объектного анализа и проектирования.

Следует здесь подчеркнуть, что появление объектных технологий и стремление к их использованию вовсе не означают полного отказа от реляционных технологий. Реляционные СУБД вместе с их развитой инструментальной инфраструктурой и созданные на их основе системы баз данных будут попрежнему активно использоваться в тех областях, где они являются достаточно эффективными.

Объектные технологии в системах баз данных уже обладают достаточно эффективными инструментальными средствами. Сформировалась быстро развивающаяся отрасль производства программных продуктов для этих целей. В последние годы создание "чисто"объектных систем значительно активизировалось. Основные их поставщики объединились в Object Data Management Group (ODMG) с целью выработки стандартов технологий объектных баз данных [42, 63]. Составляющая основу стандартов ODMG объектная модель является расширением объектной модели OMG. Благодаря этому обеспечивается естественное погружение среды ODMG в архитектуру CORBA.

Работа над стандартом ODMG 2.0 [42] и его принятие (1997) способствовали значительному расширению состава

участников этого консорциума (к их числу в настоящее время относятся такие ведущие поставщики объектных СУБД, как *O₂*, Objectivity, GemStone Software, POET Software, Versant Object Technology и др.), а также спектра поставляемых коммерческих программных продуктов для объектных систем баз данных.

Активизация объектного направления в технологиях баз данных в значительной мере стимулировалась также включением в указанный стандарт спецификаций связей для объектных языков программирования C++ и особенно Java, а также значительным прогрессом в развитии общей объектной инфраструктуры, в формировании которой, несомненно, определяющий вклад принадлежит Object Management Group (см. п. 7.5).

Несмотря на очевидные достоинства объектных технологий во многих приложениях систем баз данных, их распространение существенно сдерживается инерцией, связанной с ресурсным потенциалом, накопленным в рамках реляционных технологий, и необходимостью крупных капиталовложений для радикальной смены технологий. Нужно заметить, что эти обстоятельства существенно менее сказываются в разработках научных систем, не настолько обремененных реляционной предисторией, чем и объясняется более активное использование здесь новых объектных технологий, по сравнению с другими возможными сферами их применения.

В такой ситуации паллиативным решением стало создание гибридных *объектно-реляционных систем*, обеспечивающих в необходимых случаях для крупных реляционных систем баз данных эволюционный переход к объектным технологиям. Инструментальные средства для этой цели были созданы лидирующими производителями программного обеспечения систем баз данных - компаниями Informix, Oracle и IBM, выпустившими в 1996-1997 годах объектно-реляционные серверы баз данных. Эти СУБД, названные универсальными серверами баз данных, благодаря механизмам расширения системы типов обеспечивают поддержку новых типов не только в приложениях, но и в среде СУБД. Так, СУБД Universal Database Server (Informix), например, позволяет факультативно использовать совместно с ядром системы специальные модули DataBlade, позволяющие вводить дополнительные типы данных. Один из таких модулей, поставляемых Informix, поддерживает временной ряд как дополнительный тип данных [25] и тем самым во многом облегчает создание коллекций временных рядов и управление такими данными.

Серьезным стимулом для дальнейшего распространения объектно-реляционного подхода и базирующихся на нем технологий является ожидаемое в 1999 году принятие ANSI и ISO *стандарта SQL:1999* [17]. В новом стандарте языка SQL, наряду с развитием реляционной функциональности (новые типы данных, дополнительные предикаты, введение триггеров, усиление средств обеспечения безопасности данных и др.), предусматриваются и основные объектные возможности (структурные типы, определяемые пользователем, средства описания поведения, наследование свойств, уникальная идентификация объектов и др.). Предполагается также включить в этот стандарт поддержку темпоральных свойств данных. Конструктивная основа для соответствующих спецификаций (концепции времени, глоссарий временных баз данных, язык TSQL2) была создана еще около пяти лет назад [27, 51], и ведется работа над соответствующим разделом стандарта. В SQL:1999 будут включены также средства связи с языком Java (OLB/SQL, см. ниже).

Многие элементы нового стандарта уже реализованы в коммерческих серверах баз данных.

Что касается процессов интеграции технологий баз данных с Web-технологиями, которые мы рассматриваем как другой важный фактор, способный оказать существенное влияние на применение технологий баз данных в создании научных коллекций для ЭБ, то нужно прежде всего отметить, что эти процессы протекают в двух направлениях. С одной стороны, появляются новые версии СУБД, которые обладают дополнительными функциональными возможностями, ориентированными на использование этих систем в среде Internet. Примером могут служить СУБД Oracle8i и Informix Internet Foundation 2000, уже поставляемые компаниями-разработчиками. С другой стороны, в развитии самой среды WWW предпринимаются попытки обеспечить поддержку некоторых важных концепций, лежащих в основе методов управления данными в технологиях баз данных, что, несомненно, сближает эти области технологий. При обсуждении новых направлений в Web-технологиях мы приведем ряд конкретных иллюстраций, подтверждающих эти обстоятельства.

Наконец, нужно отметить, что *стандартизация интерфейсов* между системами баз данных и системами программирования на языке Java в значительной мере повышает роль языка Java и объектных СУБД в создании и поддержке коллекций ИР. Выше мы уже упоминали о том, что разработанные ODMG спецификации связывания для языка Java являются одним из компонентов стандарта объектных баз данных ODMG 2.0.

Аналогичная работа проводится также в области реляционных и объектно-реляционных баз данных. Индустриальный *стандарт JDBC*, разработанный компанией Javasoft, с архитектурной и функциональной точки зрения является аналогом известного стандарта ODBC и определяет спецификации интерфейса прикладного программирования на языке Java для доступа к реляционным СУБД. Первоначальная его версия входила в состав JDK 1.1. Настройкой над JDBC является *спецификация связывания SQL/OLB* (Object Language Binding) для объектных языков с базами данных SQL, уже принятая в качестве стандарта ANSI X3.135.10:1998 и представляющая собой Часть 0 стандарта SQLJ [18]. Последний, в свою очередь, планируется включить как компонент в новую версию стандарта языка SQL. Спецификации SQLJ позволяют встраивать операторы SQL в Java-программы, использовать Java в системах баз данных для реализации хранимых подпрограмм и определяемых пользователем типов данных.

7.3 Технологии документальных коллекций

В этом классе информационных технологий наиболее распространенными являются технологии *полнотекстового поиска*. Появление вычислительных систем с большой и сравнительно недорогой внешней памятью и быстрыми процессорами позволило создавать достаточно эффективные системы полнотекстового поиска. Помимо этого в настоящее время разработан и поставляется ряд реляционных и объектно-реляционных серверов баз данных, обладающих необходимыми для этого средствами. К их числу относятся, например, СУБД компаний Oracle и Informix. Тем самым появились возможности для интеграции документальных и фактографических информационных ресурсов в рамках единой базы данных.

Полнотекстовый поиск получил широкое распростра-

нение и в среде WWW благодаря созданию и предоставлению в свободный доступ мощных *поисковых серверов*. В последнее время уделяется большое внимание повышению целенаправленности поиска и селективных возможностей их поисковых механизмов. Именно этой цели служат рассмотренные выше разработки средств *описания семантики* документов в WWW [47, 48, 57, 58] (см. п. 7.1), в частности, предоставляемых средствами инфраструктуры нового языка разметки XML.

Другое важное направление развития технологий рассматриваемой категории связано с созданием *систем управления документами* (см. например [60, 61]). В настоящее время поставляются весьма разнообразные программные продукты этой категории. Наиболее развитые из них интегрируют технологии баз данных, Web-технологии, коммуникационные возможности, а также Java-технологии. Основные функции развитых систем управления документами - поддержка процесса разработки документов, их хранения, обработки, поиска и распространения с учетом возможности обслуживания мобильных пользователей, а также администрирование активными и архивными документальными ресурсами.

При этом обеспечивается не только индивидуальный, но и групповой режим разработки текстовых и мультимедийных документов с использованием методов управления потоками работ и поддержкой компонентной модели, допускается многоверсионность документов и их компонентов. В таких системах важное место занимают технологии сканирования печатных источников и распознавания оптических образов. Средства хранения обеспечивают идентификацию документов, поддерживают репозитории архивных документов, а также хранение активных документов и доступ к ним.

Системы управления документами предусматривают возможности для технической и семантической обработки документов. *Техническая обработка* включает функции просмотра и редактирования документов и их компонентов, их копирование, конвертирование в другой формат и т.д. *Семантическая обработка* - это, например, автоматическое реферирование документов, перевод их с одного языка на другой, кластеризация на основе распознавания семантики документов, представления ее в форме семантических сетей, которые далее сопоставляются с семантическими сетями заданных эталонных документов. Предусматриваются различные возможности поиска документов вплоть до полнотекстового поиска с учетом булевских условий, а также различных критериев сходства, в частности, фонетического. В распространении документов используются push-технологии и pull-технологии, электронная почта и другие средства.

На развитие возможностей документальных коллекций, безусловно, окажут влияние новые технологии семантической обработки текстовых документов, которые основаны на методологии, называемой *глубинным анализом текста* (Text Mining) [52], по аналогии с направлением Data Mining в управлении данными. Эта методология активно развивается в последнее время компанией ИВМ. Общий ее смысл состоит в обеспечении для пользователя возможностей восстановления той содержательной информации, которая была заложена в документ его создателем.

Важный этап процесса глубинного анализа связан с извлечением из текста его характерных элементов или свойств, которые могут использоваться в качестве метаданных документа, своеобразных его "дескрипторов". Сюда относятся, например, распознавание языков, на ко-

торых написан текст, извлечение упоминаемых в нем названий объектов исследований, научных учреждений или фамилий ученых. Другая важная задача, которая может быть решена с помощью извлеченных свойств, состоит в отнесении документа к некоторым категориям из заданной схемы их систематизации с одновременным индексированием для нерегламентированного поиска. Появляются также новые возможности для осуществления семантического поиска документов. Компания IBM уже успела реализовать рассматриваемые подходы в своем программном продукте Intelligent Miner for Text, предоставляющем пользователю целый комплекс инструментальных средств.

Интенсивное развитие коммуникационной среды открыло возможности для интеграции распределенных информационных ресурсов документальных систем в рамках единой коллекции. Для обеспечения семантической интероперабельности таких составных частей интегрированных коллекций необходимо решить задачу *интеграции их тезаурусов*. В настоящее время известен ряд экспериментальных проектов в этой области. Так, например, в [29] предлагается подход к решению этой задачи на основе использования медиаторов в среде архитектуры CORBA.

Следует здесь упомянуть также о новых возможностях доступа пользователей к коллекциям, обеспечивающего функции информационного поиска на основе *стандарта Z39.50* [35]. Этот стандарт специфицирует архитектуру и функции абстрактной информационной системы с развитыми возможностями доступа. Первоначально он был ориентирован, главным образом, на доступ к коллекциям текстовых документов. Действующая его версия (1995) предусматривает поддержку протоколов TCP/IP и тем самым обеспечивает возможность использования Z39.50 в среде Internet. В последнее время активно ведутся работы, цели которых состоят в расширении сферы применения этого стандарта на среду WWW, в поддержке поиска в базах данных SQL, а также интеграции его в среду CORBA-технологий [23]. Кроме того, предусматривается возможность поиска по элементам Дублинского ядра [31].

7.4 Web-технологии

Интенсивное развитие Web-технологий в последнее время связано с принятием консорциумом W3C уже упоминавшегося выше *стандарта XML* (Extensible Markup Language) [19] - спецификаций нового расширяемого языка разметки - и с развитием его инфраструктуры. Язык XML представляет собой удобное для реализации подмножество известного языка SGML (стандарт ISO).

XML - это *родовой язык*. На его основе можно конструировать разнообразные языки разметки для различных сфер применения, и эти языки имеют статус приложений XML. Одну из таких конкретизаций XML представляет язык HTML.

Как указывалось выше (см. п. 7.1) средствами XML и его инфраструктуры - ряда дополняющих его языковых спецификаций - можно задавать метаданные, описывающие структуру, содержание и другие свойства XML-документов.

После принятия рекомендаций W3C по языку XML быстро начали появляться различные использующие его инструментальные средства в статусе приложений XML, прежде всего языковые средства разметки, ориентированные на конкретные предметные области. В

контексте данной работы следует, в частности, упомянуть язык разметки математических текстов MathML (Mathematical Markup Language) [37], химический язык разметки CML (Chemical Markup Language) [40] и написанный на Java браузер, позволяющий работать с документами на CML, язык разметки для описания астрономических инструментов Astronomical Instrument Markup Language (AIML) [12]. Кроме того, как уже отмечалось выше, средствами XML DTD разработано определение типов XML-документов для коллекций исторических документов [2].

В развитии Web-технологий важное место отводится *интеграции с технологиями баз данных*. Теперь уже речь идет не о широко используемом "механическом" встраивании систем баз данных в среду WWW с доступом через HTML-формы и CGI, а о более фундаментальных подходах, предусматривающих, в том числе, и реализацию в Web-среде основных принципов управления данными, традиционных для систем баз данных. Имеется в виду создание адекватных моделей данных для Web и более общих моделей, позволяющих одновременно оперировать как базами данных, так и полуструктурированными данными Web, явная поддержка метаданных в Web для различных приложений, разработка языков запросов для Web, а также специальных архитектурных подходов, обеспечивающих интеграцию с системами баз данных и основанных, в частности, на принципах архитектуры промежуточного слоя, использовании медиаторов, адаптеров и на других известных идеях.

В последние годы в указанных направлениях ведутся весьма интенсивные исследования. Используемые в этой области подходы и инструментальные средства, их реализующие, рассматриваются, например, в обзорах [20, 36]. В [44] обсуждается разработанная в Стенфордском университете объектная модель данных Object Exchange Model (OEM), обеспечивающая представление как структур данных в базах данных, так и полуструктурированных данных Web. Рекомендации W3C [14] содержат спецификации объектной модели документов Document Object Model (DOM) для сред HTML и XML, а также описание ее отображений в язык определения интерфейсов стандарта CORBA (OMG IDL) и в язык Java.

Вопрос о средствах *спецификации метаданных* в Web-среде и об их поддержке, подробно обсуждался в п. 7.1 выше, и мы не будем здесь возвращаться к этой теме.

Предложен также целый ряд *языков запросов* для Web, основанных на различных моделях данных, в частности, языки WebSQL [5], WebOQL [4] и целый ряд других [20]. Один из языков запросов для XML разрабатывается также W3C [13]. Новые возможности в разработке коллекций открываются благодаря интеграции сред XML и систем баз данных. Уже начинают появляться программные продукты, обеспечивающие такие возможности, например, информационный сервер Tamino компании Software AG и уже упоминавшаяся новая СУБД компании Informix, предназначенная для поддержки Internet-приложений, - Informix Internet Foundation 2000.

Предпринимаются также шаги, направленные на *ит интеграцию объектных технологий OMG и Web-технологий*. Создан ряд Web-браузеров со встроенными брокерами объектных запросов, соответствующих стандарту CORBA. Разработаны спецификации протокола межброкерного обмена в Internet (ПОР), являющиеся частью стандарта CORBA [11]. Наконец, OMG принят стан-

дарт основанного на XML языка XMI [54] для обмена метаданными между средствами объектного анализа и проектирования (см. п. 7.1). Некоторые средства для рассматриваемых целей предусматриваются также в CORBA 3 - создаваемой новой версии стандарта OMG CORBA (см. п. 7.5).

В последнее время в реализации крупных проектов среда Web все активнее используется в качестве платформы для интеграции технологий интероперабельности CORBA, технологий баз данных и документальных систем, а также технологий Java.

7.5 Объектные технологии OMG для научных коллекций

Мы уже отмечали выше ту важную роль, которую играет деятельность консорциума Object Management Group (OMG) по разработке и развитию стандартов, обеспечивающих возможности создания распределенных неоднородных интероперабельных объектных сред.

Действительно, консорциумом OMG предложена архитектура CORBA для таких сред, базирующаяся на объектной модели OMG и языке определения интерфейсов IDL [11, 64], который независим от языков программирования, используемых для реализации приложений. Стандартизованы отображения спецификаций IDL в языки Ada, C, C++, COBOL, Java, Smalltalk. Создан протокол ИОП для погружения приложений CORBA в коммуникационную среду Internet. Обеспечено взаимодействие с другими широко распространенными распределенными объектными средами - DCE, COM и OLE Automation. Текущая версия этого стандарта - CORBA 2.3 [11] - принята OMG летом 1999 г.

В настоящее время в OMG началась процедура одобрения стандарта *CORBA 3* [50], ряд составных частей которого уже существует. Главные его нововведения состоят в углублении интеграции с Internet, повышении качества управления объектными сервисами и в обеспечении развитой поддержки компонентной архитектуры. Первая группа новых средств включает спецификации брандмауэра (Firewall) транспортного уровня и уровня приложений, двунаправленного соединения по протоколу CORBA GIOP, а также интероперабельного сервиса имен. Во вторую группу входят спецификации асинхронного обмена сообщениями, обеспечивающие ряд асинхронных и независимых от времени режимов вызова с возможностями статического и динамического вызова каждого из них. Рассматриваемым целям служат также три специальных конфигурации CORBA - минимальная, предназначенная, главным образом, для встроенных систем, конфигурация для систем реального времени и помехоустойчивая конфигурация. Наконец, к третьей группе относятся контейнерная среда, обеспечивающая транзакционные возможности, безопасность и сохраняемость (persistence) данных, средства интеграции с компонентной технологией JavaBeans, а также дистрибутивный формат программного обеспечения, необходимый для формирования рынка компонентного программного обеспечения CORBA.

Наряду с указанными архитектурными средствами, OMG создана также развитая инфраструктура CORBA для широкого спектра областей применения - предложены общие средства, называемые объектными сервисами, и специализированные средства для целого ряда конкретных предметных областей.

Результативна работа консорциума и по стандартиза-

ции в области *объектного анализа и проектирования*, касающаяся прежде всего средств моделирования метаданных и обмена метаданными между различными CASE-системами (см. п. 7.1). Широкое признание получил стандартизованный OMG визуальный язык моделирования UML (Unified Modeling Language) [53], предназначенный для использования в инструментальных средствах объектного анализа и проектирования программных систем.

Наряду с созданием и развитием общих технологий интероперабельности для распределенных неоднородных объектных сред, которые уже начинают находить широкое применение в нашей стране и за рубежом в создании крупных научных систем, оперирующих неоднородными распределенными коллекциями информационных ресурсов (см. например [62, 66, 67]), OMG в последние годы предпринимает усилия по созданию специализированного стандартизованного инструментария такого рода для поддержки научных исследований.

Так, с 1997 г. в рамках деятельности OMG по развитию средств архитектуры CORBA для различных конкретных сфер применения ("вертикальный рынок") ведется разработка комплекса объектных стандартов для поддержки исследований в ряде областей *наук о жизни* (Life Sciences Research, LSR), в том числе, в биоинформатике (в частности, в исследованиях геномов и структурной биологии), химической информатике, медицинских клинических испытаниях, вычислительной химии [33].

Предполагается прежде всего сформировать основанную на технологии CORBA единую общую архитектуру приложений для проведения исследований в указанных областях [15], и для ее реализации разработать комплекс стандартов OMG (см. например [6, 8, 10, 21]). Некоторые из них планируется принять уже в 1999 г. Такой инструментарий может быть использован для создания и поддержки научных объектных коллекций на основе архитектуры CORBA в указанных областях исследований.

Представляет также интерес разработка спецификаций OMG для набора обобщенных *библиографических сервисов*, основанных на архитектуре CORBA, которые обеспечат доступ к неоднородным библиографическим базам данных и разработку клиентских средств для их использования [34]. При этом предусматривается возможность работы с библиографическими ссылками не только на публикации традиционной природы (статьи, книги, диссертации и т.п.), но и на записи в базах данных, электронные издания, Web-сайты, мультимедийные ИР. Выработку предложений по этой проблеме предполагается завершить в начале 2000 г. Хотя указанная работа выполняется с ориентацией на LSR, она имеет, несомненно, общенаучный интерес.

8 Заключение

Активное развитие работ в нашей стране по созданию электронных библиотек и коллекций информационных ресурсов для таких систем, несомненно, будет способствовать созданию эффективной инфраструктуры для поддержки научных исследований и других сфер деятельности.

Актуальное значение имеет разработка новых информационных технологий, не только интегрирующих известные достижения, но и основанных на оригинальных подходах. Вместе с тем, важнейшей задачей в рассматриваемой области является полноценное использование возможностей существующих перспективных технологий для практической реализации электронных библиотек.

Необходимо также преодолеть соблазн попыток воспользоваться появившимся модным флагом при проведении разработок, основанных на традиционных технологиях, без осуществления реальных шагов к достижению качественно нового уровня функциональности.

Библиография

- [1] *A proposed convention for embedding metadata in HTML*. A position paper from May 1996, W3C Workshop on Distributed Indexing and Searching. [<http://www.w3.org/Search/9605-Indexing-Workshop/ReportOutcomes/S6Group2.html>]
- [2] *American Memory DTD for Historical Documents*. [<http://lcweb2.loc.gov/ammem/amdtd.html>]
- [3] Arms C.R. *Historical Collections for the National Digital Library*. D-Lib Magazine, April 1996.
- [4] Arocena G., Mendelson A. *Viewing Web Information Systems as Database Applications*. Comm. of the ACM, July 1998.
- [5] Arocena G.O., Mendelson A.O., Mihaila G.A. *Applications of a Web Query Language*. Department of Computer Science, University of Toronto, 1996. [<http://www.cs.toronto.edu/~websql/www-conf/wsqli/PAPER267.html>]
- [6] *Biomolecular Sequence Analysis*. RFP Response. Initial Submission. OMG Document lifesci/98-10-04.
- [7] *CASE Data Interchange Format (CDIF) - Overview*. Electronic Industries Association. CDIF Technical Committee. January 1994.
- [8] *Chemical Entity Representation and Interface Definition. Request for Information*. OMG Document lifesci/99-03-05.
- [9] Chen P.P. *The entity-relationship model – toward a unified view of data*. ACM TODS, 1(1): 9-36, March 1976.
- [10] *Clinical Trials Workgroup White Paper, Draft B*. OMG Document lifesci/98-06-01.
- [11] *Common Object Request Broker Architecture*. Version 2.3. Object Management Group, June 1999. OMG Documents formal/99-07-01 - formal/99-07-28.
- [12] Cover R. *Astronomical Instrument Markup Language (AIML)* [<http://www.oasis-open.org/cover/aiml.html>]
- [13] Deutsch A., Fernandez M., Florescu D., Levy A., Suci D. *XML-QL: A Query Language for XML*. Submission to the WWWC, August 19, 1998. [<http://www.w3.org/TR/1998/NOTE-xml-ql-19980819>]
- [14] *Document Object Model (DOM) Level 1 Specification*. Version 1.0. W3C Recommendation. October 1, 1998. REC-DOM-Level-1-19981001. [<http://www.w3.org/TR/REC-DOM-Level-1>]
- [15] *Domain Software Architecture for Life Sciences Research*. OMG Document lifesci/99-03-07.
- [16] *Dublin Core Metadata Element Set Reference Description*, Version 1.1, 1999-07-02. [http://purl.org/dc/documents/proposed_recommendations/pr-dces-19990702.htm]
- [17] Eisenberg A., Melton J. *SQL:1999, formerly known as SQL3*. SIGMOD Record, Vol. 28, No. 1, March 1999. Есть русск. перевод: Эйзенберг Э., Мелтон Д. *SQL:1999, ранее известный как SQL3*. Открытые системы, 1, 1999.
- [18] Eisenberg A., Melton J. *SQLJ Part 0, now known as SQL/OLB (Object Language Bindings)*. SIGMOD Record, Vol. 27, No. 4, December 1998. Есть русск. пер.: Эйзенберг Э., Мелтон Дж. *Связывания для объектных языков: SQLJ Часть 0, называемая теперь SQL/OLB*. Открытые системы, 4, 1999.
- [19] *Extensible Markup Language (XML) 1.0*. W3C Recommendation 10-February-1998 [<http://www.w3.org/TR/1998/REC-xml-19980210>]
- [20] Florescu D., Levy A., Mendelson A. *Database Techniques for the World-Wide Web: A Survey*. SIGMOD Record, Vol. 27, No. 3, September 1998. Есть русск. пер.: Флореску Д., Леви А., Мендельсон А. *Технологии баз данных для World-Wide Web: обзор*. СУБД, 4-5/1998.
- [21] *Genomic Maps RFP*. OMG Document lifesci/98-11-07.
- [22] Griffin S.M. *NSF/DARPA/NASA Digital Libraries Initiative*. D-Lib Magazine, July/August 1998.
- [23] Hammer S., Favaro J. *Z39.50 and World Wide Web*. D-Lib Magazine, March 1996.
- [24] *HTML 4.0 Specification*. W3C Recommendation, Revised on 24-Apr-1998. [<http://www.w3.org/TR/REC-html40>]
- [25] *Informix TimeSeries DataBlade Module. User's Guide*. Version 3.1. Informix Software Inc. April 1998.
- [26] *ISO/IEC 10027:1990 Information Resource Dictionary System (IRDS) Framework*.
- [27] Jensen C.S., ed. *A consensus Glossary of Temporal Database Concepts*. SIGMOD Record, Vol. 23, No. 1, March 1994.
- [28] Kogalovsky M.R. *Time Series Relation Data Model*. Proc. of the International Workshop on Advances in Databases and Information Systems - ADBIS'94, Institute for Problems of Informatics, Russian Academy of Sciences, Moscow, 1994.
- [29] Kramer R., Nikolai R., Habeck C. *Thesaurus federations: loosely integrated thesauri for document retrieval in networks based on Internet technologies*. Intern. Journal on Digital Library, 1, 1997.
- [30] Lagoze C., Fielding D. *Defining Collections in Distributed Digital Libraries*. D-Lib Magazine, November 1998.
- [31] LeVan R. *Dublin Core and Z39.50. OCLC Office of Research and Special Project*. Draft Version 1.2, February 1998. [<http://cypress.dev.oclc.org:12345/~rll/docs/dublincoreandz3950.html>]

- [32] Ley M. *Computer Science Bibliography*. Universitat Trier. [<http://dblp.uni-trier.de/>]
- [33] *Life Sciences Research Domain Task Force Pocket Guide*. OMG Document lifesci/99-04-02.
- [34] *LSR Bibliographic Query Services. Request For Proposal. Draft D*. Document OMG lifesci/99-03-10.
- [35] Lynch C.A. *The Z39.50 Information Retrieval Standard*. D-Lib Magazine, April 1997.
- [36] Manola F. *Toward a Web Object Model*. Object Services and Consulting, Inc. February 10, 1998. [<http://www.objs.com/O5A/wom.htm>]
- [37] *Mathematical Markup Language (MathML) 1.0 Specification*. World Wide Web Center Recommendation 07-April-1998. [<http://www.w3.org/TR/1998/RECMathML-19980407>]
- [38] *Meta Object Facility (MOF) Specification*. October 7, 1997. Joint Revised Submission. OMG Document ad/97-10-02.
- [39] *Metadata Interchange Specification (MDIS)*. Version 1.1. Meta Data Coalition, August 1, 1997.
- [40] Murray-Rust P. *JUMBO and XML/CML demonstration*. [<http://www.nottingham.ac.uk/~pazpmr/README/>]
- [41] *Namespaces in XML*. World Wide Web Consortium 14-January-1999. REC-xml-names-19990114.
<http://www.w3.org/TR/1999/REC-xml-names-19990114>
- [42] *Object Database Standard: ODMG 2.0*. Ed. by R.G.G. Cattell, D.K. Barry. Morgan Kaufmann Publishers, Inc. 1997.
- [43] *Open Information Model (OIS)*, Version 1.0. Review Draft. Meta Data Coalition. April 12, 1999.
- [44] Papakonstantinou Y., Garsia-Molina H., and Widom J. *Object Exchange Across Heterogeneous Information Sources*. IEEE Int. Conf. on Data Engineering, Taipei, March 1995.
- [45] *Report on NSF Workshop "Distributed Knowledge Work Environment"*. March 9-11, 1997. Santa Fe, New Mexico.
<http://www.si.umich.edu/SantaFe/>
- [46] *Report on Workshop "Interoperability, Scaling, and Digital Library Research Agenda"*, 1995. [<http://www.ccc.gov/pubs/iita-dlw/>]
- [47] *Resource Description Framework (RDF). Model and Syntax Specification*. W3C Recommendation 22 February 1999.
<http://www.w3.org/TR/REC-rdf-syntax/>
- [48] *Resource Description Framework (RDF). Schema Specification*. W3C Proposed Recommendation 03 March 1999
<http://www.w3.org/TR/PR-rdf-schema/>
- [49] Scherlis W.L. "Foreground" information stores, or personal digital libraries. To the Editor. What's needed in future research? D-Lib Magazine, July 1995.
- [50] Sigel J. *What's Coming in CORBA 3* [<http://www.omg.org/news/pr98/component.html>]
- [51] Snodgrass R.T., Ahn I., Ariav G., Batory D.S., Clifford J., Dyreson C.E., Elmasri R., Grandi F., Jensen C.S., Kafer W., Kline N., Kulkarni K.G., Leung T.Y.C., Lorentzos N.A., Roddick J.F., Segev A., Soo M.D., and Sripada S.M. *A TSQL2 Tutorial*. SIGMOD Record, Vol. 23, No. 3, 1994.
- [52] *Text Mining Technology. Turning Information Into Knowledge. A White Paper from IBM*. Ed. by D. Tkach. IBM Software Solution. February 17, 1998.
- [53] *UML Specification*. OMG Documents ad/97-08-02 - ad/97-08-09.
- [54] *XML Metadata Interchange (XMI)*. Proposal to the OMG OA&DTF RFP3: Stream-based Model Interchange Format (SMIF). OMG Document ad/98-10-05.
- [55] *XML Schema Part 1: Structures*. W3C Working Draft 6-May-1999 [<http://www.w3.org/TR/xmlschema-1>]
- [56] *XML Schema Part 2: Datatypes*. W3C Working Draft, 6-May-1999 [<http://www.w3.org/TR/xmlschema-2>]
- [57] Weibel S. *Metadata: The Foundations of Resource Description*. D-Lib Magazine, July 1995.
- [58] Weibel S. *The State of the Dublin Core Metadata Initiative*. April 1999. D-Lib Magazine, April 1999.
- [59] *What is OMG-UML and why is it important?* Object Management Group Press Release, 1997. Есть русск. пер.: *Что такое OMG-UML и почему он важен*. Открытые системы, 1, 1999.
- [60] Баласанян В. *Концепции системы автоматизации отечественного документооборота*. Открытые системы, 1, 1997.
- [61] Гавердовский А. *Концепции построения систем автоматизации документооборота*. Открытые системы, 1, 1997.
- [62] Захаров В.Н. *Создание интегрированных электронных библиотек на основе неоднородных, распределенных электронных коллекций научной информации*. Институт проблем информатики РАН. Проект РФФИ 98-07-91061.
- [63] Калининченко Л.А. *Стандарт систем управления объектными базами данных ODMG-93: краткий обзор и оценка состояния*. СУБД, 1/1996.
- [64] Калининченко Л.А., Коголовский М.Р. *Стандарты OMG: Язык определения интерфейсов IDL в архитектуре CORBA*. СУБД, 2/1996.
- [65] Коголовский М.Р. *Базы данных в экономико-математическом моделировании: методология, приложения, инструментарий* //Труды семинара Московской секции ACM SIGMOD "Перспективы развития систем баз данных и информационных систем" - ADBIS'93, Москва, май 1993. - М.: Ин-т проблем информатики РАН, 1993.

- [66] Колчанов Н.А. *Интегральная электронная библиотека по пространственным структурам и функциям ДНК, РНК и белков (в составе Электронной библиотеки СО РАН)*. Институт цитологии и генетики СО РАН. Проект РФФИ 98-07-91078.
- [67] Марчук А.Г., Осипов А.Е. *Обеспечение унифицированного доступа к разнородным коллекциям и информационным ресурсам на основе технологии CORBA*. ИСИ СО РАН. Проект РФФИ 98-07-91256.