СИСТЕМАТИКА КОЛЛЕКЦИЙ ИНФОРМАЦИОННЫХ РЕСУРСОВ В ЭЛЕКТРОННЫХ БИБЛИОТЕКАХ

М.Р. Когаловский
Институт проблем рынка РАН
e-mail: kogalov@cemi.rssi.ru

Аннотация

Одним из важных направлений активно развивающихся в последние годы разработок систем электронных библиотек, является создание для них коллекций информационных ресурсов. Коллекции 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 коллекции обладают и некоторыми общими свойствами, учитывать которые необходимо при их разработке.

Предлагаемая работа посвящена вопросам методологии разработки коллекций. В ней обсуждаются основные общие свойства коллекций информационных ресурсов, методы систематизации, применяемые при их формировании, рассматриваются вопросы генезиса коллекций, роль в них метаданных, 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 Коллекции информационных ресурсов

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

С точки зрения уровня абстракции, ИР коллекции подразделяются на *данные* (информацию) и *метаданные* (метаинформацию). Ресурсы первого вида представляют интересующие пользователей сведения о предметной области этой коллекции. В свою очередь, метаданные коллекции характеризуют свойства самой коллекции в целом и принадлежащих ей ресурсов как сущностей реального мира, используемые средствами управления коллекцией и/или пользователями для корректной интерпретации ИР. 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, для некоторых приложений 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). Адекватная систематизация ИР не только облегчает доступ пользователей к ним, но и дает возможность целенаправленно и рациональным образом исследовать с их помощью предметную область коллекции.

2 Свойства коллекций

Свойства коллекций, играющие существенную роль в управлении ими и их использовании, можно разделить на несколько категорий - *идентифицирующие свойства* (название и/или другие идентификаторы коллекции), *свойства содержания* (общее описание содержания коллекции какими-либо средствами, например, аннотация, ключевые слова и т.д.; состав информационных ресурсов, явная/неявная его спецификация; связи данной коллекции с другими коллекциями; степень ее полноты; оценки непротиворечивости, достоверности информационных ресурсов; ограничения целостности), *прагматические свойства* (назначение; область использования; социальная значимость; круг пользователей; условия доступа; допустимые гео-

графические, административные и/или временные рамки применения), *характеристики генезиса* (источники, методы и процессы создания; история коллекции), *организационные свойства* (способ задания состава коллекции; методы систематизации ее информационных ресурсов; принципы их именования; структура; характеристики объема и динамики; местонахождение), *технологические свойства* (базовые технологии; стандарты; методология и инструментарий разработки; инструментальные средства поддержки; архитектура; представление - среда, актуальная/виртуальная коллекция, однородные/неоднородные ресурсы, связь с источниками, формат, кодирование, единицы измерения значений, язык или языки представления; способы, протоколы и средства доступа), *правовые свойства* (авторские права; права владения, пользования; издательские права).

Кратко обсудим некоторые из перечисленных свойств. Методы систематизации коллекций, вопросы их формирования и генезиса, роль метаданных в коллекциях, особенности научных коллекций, информационные 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 ресурсов WWW). Другой подход предусматривает *неявное* задание состава ИР путем спецификации в какой-либо форме *критерия принадлежности* ИР данной коллекции (Membership Criteria) [34]. 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. Статические коллекции информационных ресурсов часто распространяются на CD-ROM. Коллекциями такого рода являются CD-версии ресурсов многочисленных электронных изданий и некоторых электронных библиотек, например, Антологии ACM SIGMOD [2].

Во многих коллекциях одним из элементов их систематизации является *единая схема именования* ИР, принадлежащих коллекции. Выбор эффективных правил именования ИР весьма существенен, особенно для крупных коллекций. Лаконичные имена, легко ассоциируемые с обозначаемыми ими ресурсами коллекции, эффективно выполняют функцию идентификации ИР, существенно упрощают работу пользователей. Обязательное ограничение, налагаемое на схему именования, заключается в том, что присваиваемые ресурсам имена должны быть *уникальными* в рамках коллекции в целом или некоторых подмножеств ее ИР. Иногда на имя ресурса наряду с функцией идентификации возлагается роль указателя (логического) места ИР 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 моносреде - тексты, изображения, аудио или видео - или являются мультимедийными. Организация и способы представления ИР коллекции, в том числе, и метаданных, существенным образом зависят от информационных технологий, используемых в данной ЭБ (см. п. 8).

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 коллекциях слабоструктурированных данных Web, представленных средствами

языка HTML, не обеспечивается какой-либо автоматики, предназначенной для этих целей. Положение для коллекций в среде Web-технологий изменяется с появлением нового языка разметки XML и его инфраструктуры [23, 64].

3 Свойства информационных ресурсов

Что касается свойств отдельного информационного ресурса, то он наследует некоторые свойства коллекции в целом и, кроме того, может обладать собственными индивидуальными свойствами. В последние годы предпринимаются попытки стандартизации наборов свойств (атрибутов), которые можно было бы использовать для описания полнотекстовых документов, поддерживаемых в среде WWW, и тем самым повысить эффективность поиска таких ресурсов. С этой целью был разработан и продолжает развиваться стандарт, названный *Дублинским ядром* [19] (см. подробнее об этом в п. 8.1). Он начинает применяться для описания свойств не только полнотекстовых документов, но и информационных ресурсов различных других видов.

4 Методы систематизации коллекций

Систематизация коллекции осуществляется, как правило, на основе свойств ее предметной области (ПО), свойств ИР коллекции, пользователей, процессов формирования коллекции, участников этих процессов, по хронологическому принципу и в других аспектах.

Для систематизация по свойствам ПО используется *концептуальная модель ПО*, построенная в результате ее изучения и анализа. Важно заметить, что с некоторыми коллекциями, например, с научными коллекциями экспериментальных данных, может ассоциироваться несколько сосуществующих концептуальных моделей ПО, соответствующих различным теориям исследуемого феномена (см., например [76]). В таких случаях использование разных моделей приводит к различным интерпретациям содержания одних и тех же ИР коллекции и к разным вариантам их систематизации.

Концептуальная модель ПО может быть представлена различными средствами. При реализации коллекции на основе технологии баз данных эта модель отображается в среду конкретной выбранной инструментальной СУБД и описывается *концептуальной схемой базы данных* [71]. 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.

Классы ИР коллекции в процессе их систематизации на основе концептуальной модели ПО ассоциируются чаще всего с типами сущностей или связей, с экземплярами (или множествами экземпляров) сущностей, обладающих заданным набором свойств и т.д.

Заметим, что модели данных, поддерживаемые коммерческими СУБД общего назначения не всегда являются приемлемым инструментом для этих целей. Так, широко распространенные реляционные СУБД не обладают рядом функциональных возможностей, необходимых для создания и поддержки в их среде научных коллекций ИР. Они не поддерживают сложные типы и структуры данных, темпоральные и пространственные свойства данных, их многоверсионность и т.д. По этим причинам для создания и поддержки коллекций ИР средствами баз данных в ряде случаев приходится создавать *специальные модели данных* и реализовать их средствами более высокий уровень представления данных над реляционной моделью. Специализированные модели данных часто строятся как расширения реляционной модели. Этот подход был использован автором при разработке инструментария для создания и поддержки коллекций временных рядов экономических показателей [32, 72].

Для систематизации коллекций ресурсов Web могут использоваться модели слабоструктурированых данных, которые позволяют специальным информационным серверам управлять такими коллекциями. Для коллекций неоднородных ИР используются различные интегрирующие модели.

Для многих коллекций в качестве основы систематизации ИР используют *классификаторы*, определяющие одномерные или многомерные пространства классификационных признаков. В качестве классификационных признаков выбираются при этом какие-либо из указанных выше аспектов систематизации коллекции. Если это - свойства ПО коллекции, то такой классификатор можно рассматривать как простейшую концептуальную модель ПО. Широко известным примером классификатора является УДК.

Как правило, множество возможных значений классификационных признаков может быть предопределено. При этом множество значений отдельного признака может иметь линейную или иерархическую структуру. Классы ИР коллекции соотносятся с отдельными (не обязательно со всеми) точками или гиперплоскостями пространства классификационных признаков.

Основные принципы построения классификаторов заключаются в *единстве критерия классификации* (основания классификации), *независимости* значений классификационных признаков, благодаря чему они устанавливают на множестве классифицируемых ИР отношение эквивалентности (разбиение ИР на попарнонепересекающиеся классы), в *полноте* множеств значений признаков.

Многомерные линейные и иерархические классификаторы широко применяются, в частности, для систематизации коллекций временных рядов макроэкономических показателей (имеются, 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 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. Позиционное кодирование позволяет отражать в значениях кодов иерархические отношения между значениями признаков, обеспечивая простоту процедур агрегирования данных в статистических коллекциях.

5 Генезис и формирование коллекций

В процессе формирования коллекции необходимо решить ряд взаимосвязанных задач. Основные из них - определение содержательного состава и принципов систематизации коллекции, исходя из ее назначения, выбор источников включаемых в нее ИР, обеспечение их полноты и непротиворечивости, оценка и выбор информационных технологий для формирования, поддержки и использования коллекции.

Для формирования коллекций используются разнообразные источники ИР: 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 ИР и т.д.

В некоторых случаях ИР, включаемые в формируемую коллекцию, уже существуют в оцифрованном 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 указанных технологий. Тогда имеется два подхода. При первом из них новая коллекция может создаваться как *материализованное* собрание ресурсов, используемое автономно от их источников. При этом возникают, как правило, проблемы поиска и выборки требуемых ИР из источников (например, из каких-либо коллекций), 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. Состояние такой коллекции является зависимым от состояния источников, а ее состав может быть задан *неявным* образом (см. п. 2) либо как список ссылок на источники составляющих ее ИР. Случай неявного задания обладает рядом достоинств, но могут возникнуть достаточно сложные проблемы интеграции ресурсов, 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 и различные методы "очистки" данных (data scrubbing, data cleansing).

Если коллекция формируется в течение продолжительного времени и/или может включать ИР, относящиеся к различным временным периодам существования ПО, то в силу происходящих в ПО изменений или появления новых знаний о ней могут потребоваться существенные изменения в систематизации коллекции (см. п. 6).

Во многих случаях включаемые в коллекцию ИР нуждаются в предварительной *оцифровке*. Таковы, например, ретроспективные журнальные публикации. Как правило, для этих целей могут использоваться известные типовые технологии. Однако иногда требуется специальное оборудование и программное обеспечение. Так, при автоматической регистрации показаний некоторых научных приборов необходимы специальные интерфейсные средства для их сопряжения со средствами вычислительной техники, использующие преобразователи аналоговых сигналов в цифровые, а также специализированное программное обеспечение.

Информационные технологии, используемые для создания и поддержки коллекций ИР, подробно рассматриваются в п. 8.

6 Метаданные в коллекциях

Метаданные коллекции - это совокупность значений некоторых ее свойств и свойств принадлежащих ей информационных ресурсов (п. 2,3). Конкретные функции метаданных и их состав могут значительно различаться, в зависимости от специфики ЭБ и конкретной коллекции. Действительно, для ИР в коллекциях экспериментальных данных метаданные должны указывать 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.

Как уже отмечалось, различаются *системные* и *пользовательские* метаданные. Первые из них обеспечивают функционирование механизмов управления коллекцией, а вторые описывают различные свойства коллекции и ее ИР в воспринимаемой пользователем форме.

Степень формализованности представления метаданных зависит от характера их использования - системные метаданные обычно представляются в формализованном виде, а пользовательские чаще всего на естественном языке.

С метаданными коллекций связана достаточно сложная проблема, уже упоминавшаяся в п. 5 и заключающаяся в следующем. В период создания и использования коллекции в ее предметной области и/или в свойствах коллекции могут происходить такие изменения, которые приводят к необходимости соответствующих изменений в систематизации коллекции, например, изменений набора значений классификационных признаков. Более того, может возникнуть необходимость в радикальном изменении положенной в основу коллекции классификационной схемы.

Такие ситуации исследовались в области систем баз данных. Они связаны с темпоральными свойствами данных и/или метаданных и могут приводить к эволюции 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.

Типичным примером коллекций, для которых такие явления являются правилом, а не исключением, могут служить коллекции экспериментальных данных в различных областях науки, где должна обеспечиваться сопоставимость результатов экспериментов, проводившихся в различное время и при изменяющихся условиях. Другой пример - коллекции временных рядов экономических показателей. В них по необходимости должны находить отражение структурные сдвиги в экономике, изменения хозяйственного механизма, последствия 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. Причем, 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 ряде международных, национальных и индустриальных стандартов. Основные из них будут рассмотрены далее в п. 8.1.

7 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 для них.

Научные коллекции различаются, прежде всего, *масштабом круга пользователей*. Имеют право на жизнь как коллекции, имеющие национальную или мировую значимость и предоставляемые для глобального доступа, так и персональные коллекции отдельных ученых и коллекции коллективов исследовательских лабораторий. Важно заметить, что эта характеристика коллекции может изменяться с течением времени. Персональная коллекция может со време-

нем приобрести высокий уровень значимости для многих ученых, и потребуется 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 для обеспечения доступа к коллекциям должны предусматриваться весьма гибкие пользовательские интерфейсы, способные адаптироваться к изменению потребностей пользователей.

В отличие от других видов коллекций, ИР научных коллекций могут представлять 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* (и, возможно, даже противоречивые) сведения об исследуемых объектах, явлениях или процессах. В таких ситуациях, конечно, не может идти речь об интегральной целостности данных в коллекции.

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

При разработке коллекций ИР в естественнонаучных исследованиях (например, в механике, теоретической физике) для одной и той же коллекции может сосуществовать *несколько гипотетических концептуальных моделей* предметной области, в соответствии с различными предложенными теориями и моделями исследуемого феномена [76]. Это позволяет, в частности, различным образом интерпретировать данные одних и тех же экспериментов.

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 несколькими 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 экспериментов и их результаты. Здесь важно также обеспечить средства предоставления ИР 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. Основной вид ИР таких коллекций - это временные и/или пространственные ряды наблюдений. При разработке таких коллекций и механизмов доступа 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 [5], 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. Примером такого подхода может служить общедоступная в среде WWW библиография по системам баз данных и логическому программированию [36], охватывающая публикации в многочисленных журналах и трудах конференций, а также известные 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 ACM и IEEE, а также ряда издательств, если он обладает необходимыми полномочиями.

8 Технологии и стандарты

В разработке коллекций ИР для ЭБ наряду с традиционными технологиями баз данных, текстовых систем и Web-технологиями, находят применение новые подходы, формирующиеся в каждом из этих направлений, а также смешанные интегрированные технологии. В научных коллекциях начинают использоваться некоторые ранее созданные технологии, например, технологии интероперабельности CORBA [14], которые уже достаточно активно применяются в разработках крупных корпоративных информационных систем и в других областях. Важное значение для коллекций ИР имеют технологии управления метаданными.

В этой работе не преследуется цель детального и всестороннего обсуждения технологических аспектов разработки коллекций. Мы ограничимся краткой оценкой состояния и определяющих факторов развития важнейших технологий создания, поддержки и применения коллекций, основных стандартов, на которых они базируются, а также некоторого нового инструментария для научных коллекций. При этом принимаются во внимание лишь технологии промышленного характера либо близкие к достижению такого статуса.

8.1 Управление метаданными

Как уже отмечалось, в настоящее время разработаны и используются разнообразные подходы к управлению метаданными информационных систем, принят ряд международных, национальных и индустриальных стандартов в этой области, многие из которых могут найти применение в коллекциях ИР электронных библиотек. Рассмотрим кратко наиболее перспективные из них.

При использовании техники баз данных структурные метаданные и метаданные, описывающие ограничения целостности данных, представляются в схеме базы данных, специфицируемой средствами *языка описания данных* СУБД. Стандартизованное представление метаданных в реляционных базах данных обеспечивается подмножеством языка SQL, называемым *информационной схемой*.

Для более интенсивной поддержки метаданных в системах баз данных и других типах информационных систем, разрабатываемых с помощью инструментальных средств анализа и проектирования, создан поддерживаемый ISO/IEC JTC1 стандарт *Information Resource Dictionary System (IRDS)* [30]. Он описывает системы, предназначенные 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 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 [14], созданы спецификации средств экспорта/импорта для IRDS, обеспечивается поддержка возможностей именования ресурсов и тезаурусов. Стандарт IRDS используется для создания репозиториев метаданных в системах баз данных и хранилищах данных. По отношению к коллекциям ИР в ЭБ он может рассматриваться как средство нижнего 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)* [42]. Этот стандарт основан на объектной модели, базирующейся на концепциях известной модели "сущностей-связей" П. Че-

на [11] 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 [14, 68], а также спецификации основанных на CORBA сервисов для управления метаинформацией. Модели MOF отводится роль мета-мета модели для описания мета-моделей, которые лежат в основе различных средств объектного анализа и проектирования. Авторами стандарта 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 коллекций может стать язык *UML (Unified Modeling Language)* [56, 62]. Этот язык был принят OMG одновременно с MOF в сентябре 1997 г. в качестве индустриального стандарта, призванного обеспечивать интероперабельность объектно-ориентированных инструментальных средств анализа и проектирования систем, опирающихся на архитектуру CORBA.

UML создан известными специалистами в области объектного анализа и проектирования Г. Бучем, И. Якобсоном и Д. Рамбо (компания Rational Software). В языке синтезируются и развиваются ранее разработанные ими подходы и методы (метод Буча, OMT, OOSE), учитываются возможности других получивших широкое признание методологий. Он является независимым от конкретных языков программирования, используемых при реализации проектируемых систем, 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).

Серьезно озабочено необходимостью явного использования метаданных сообщество разработчиков Web-технологий. Это позволило бы повысить эффективность поиска ресурсов поисковыми машинами Web, обеспечить основу для решения проблем семантической интеграции распределенных информационных ресурсов и их повторного использования.

Ранние попытки, предпринятые в этом направлении, привели к включению в версию языка *HTML 2.0* простейших средств, которые позволили 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 [61].

Указанные средства языка HTML получили дальнейшее развитие в версии *HTML 4.0* [28] 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 (см. ниже). В частности, для тега META были введены дополнительные атрибуты LANG и SCHEMA, позволяющие задать, соответственно, язык представления значения элемента метаданных в этом теге и уточняющий контекст для адекватной его интерпретации. Добавлен новый атрибут профиля документа PROFILE в тег HEAD заголовка, содержащий ссылку (URL) на ресурс Web, в котором определяются элементы метаданных документа и их значения. Формат содержания профиля в языке не регламентирован.

Проблема разработки средств представления метаданных возникла и в связи с созданием новых Web-технологий, основанных на языке разметки XML [23] - стандарте консорциума W3C. Прежде всего, некоторые возможности для этой цели предусмотрены в самом языке XML. Их называют декларацией типа документов. Спецификация этих деклараций средствами XML называется *Document Type Definition (DTD)* и по-

зволяет описывать допустимые структуры гипертекстовых документов рассматриваемой категории в терминах составляющих их элементов.

Для каждого типа элементов документа указывается вид их содержания (содержание отсутствует, литерная строка, список вложенных элементов, смешанное содержание - литерная строка и вложенные элементы), обязательны или факультативны данные элементы в документе, а также список 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 [3], для описания астрономических инструментов [15] 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 [58, 59].

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.

Поскольку язык XML Schema является приложением XML, то схема, специфицированная его средствами, 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 [26], консорциум W3C принял также стандарт средств для описания семантики ИР в среде Web, независимых от конкретной предметной области, - *Resource Definition Framework (RDF)*. Этот стандарт состоит из двух частей. В первой из них [51] предлагается семантическая модель и синтаксис основанного на XML языка для представления семантики ИР - RDF-спецификации. Использование здесь XML как базовой языковой среды естественным образом решает проблему обмена метаданными в WWW и их повторного использования в приложениях, основанных на XML.

Описание семантики ИР в терминах модели RDF по существу эквивалентно *ER-диаграмме* [11] и декларирует множество ИР, 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. ИР идентифицируются уникальным образом с помощью их URI (Uniform Resource Identifier, обобщение концепции URL в WWW). Они могут также представлять собой коллекции других ИР или литералов, называемые контейнерами. Допускаются контейнеры типа мультимножества, последовательности и альтернативы.

Для того, чтобы RDF-спецификация семантики ИР была полной, необходимо ассоциировать с нею описание семантики используемых в этой 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 ИР, которые описываются данной RDF-спецификацией и этой схемой.

В стандарте RDF предусматривается два способа задания схем. Первый из них, более простой, использует в качестве схемы пространство имен свойств XML - *XML-Namespace*. Спецификации пространства имен [46] - 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-спецификациях, 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 являют-

ся 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 [52] - 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 ИР, 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 [51]. 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. В этой модели используются концепции классов, свойств и ограничений, ассоциируемых с классами и свойствами, поддерживается иерархическое отношение класс-подкласс. Базовая модель в RDF Schema служит фактически метамоделью по отношению к модели, лежащей в основе RDF-спецификаций. Средствами этой модели в стандарте [52] определяется *схема-ядро*, в терминах которой описываются конкретные схемы. Схема-ядро неявно включается в состав каждой из них. Ее ресурсы составляет небольшой набор "встроенных" классов, свойств и ограничений целостности.

Метаданные, представленные средствами RDF, могут использоваться для более эффективного поиска ресурсов поисковыми машинами Web, в электронных библиотеках, в описаниях коллекций страниц Web, составляющих некоторый виртуальный документ, для представления содержания ИР 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 в разных случаях. Для некоторых приложений уже существуют такого рода стандарты. Например, для описания семантики электронных текстовых документов предложен набор свойств, названный Дублинским ядром [60]. В стандарте RDF показано, каким образом Дублинское ядро может быть выражено средствами RDF Schema [52].

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 ИР, представленных в WWW, и тем самым обеспечивали бы более эффективный поиск требуемых ресурсов, а также поддержку других Web-приложений. Выработанный на симпозиуме подход стал называться *Дублинской инициативой* [60].

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 [19] 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 продуцирован 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 [52] может быть специфицирова-

на 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 [61]. 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 коллекций, большое значение имеет стандартизация спецификаций *обмена метаданными* между различными инструментальными средствами разработки. Необходимость такого обмена связана с обеспечением повторного использования ИР коллекций, а также с решением задач реинженерии использующих их приложений.

Одна из ранних попыток в этом направлении привела еще в 1987 г. к разработке Electronic Industries Association (EIA) стандарта спецификаций *CASE Data Interchange Format (CDIF)* [9] обмена метаданными между инструментальными средствами CASE. В этом проекте использовалась специально созданная метамодель, основанная на парадигме моделирования "сущностей-связей" [11]. 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.

Автором другого стандарта аналогичного назначения стал образованный в 1995 г. консорциум *Meta Data Coalition (MDC)*, в состав которого входят такие крупные компании, как Informix, Sybase, SAS, Platinum Technology, SAS и др. Первая версия стандарта *Meta Data Interchange Specification (MDIS)* была принята в 1996 г. [43], и MDC продолжает его развивать. После вступления в MDC компании Microsoft в конце 1998 г. на основе ее предложений был создан новый стандарт спецификаций метаданных - *Open Information Model (OIS)* [49]. В апреле 1999 г. было объявлено об объединении усилий MDC и OMG для развития стандартов метаданных. В настоящее время ведутся работы по созданию формата кодирования спецификаций OIS (описываемых с помощью диаграмм UML) средствами технологий XML. Возможно, для этой цели будет использоваться новый язык OMG XMI.

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)* [57]. Назначение XMI состоит в обеспечении простого обмена метаданными между инструментальными средствами моделирования, поддерживающими язык UML [56, 62], 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 активно проводятся работы по стандартизации управления метаданными в части содержания поддерживаемых метаданных, методов их представления, а также обмена метаданными между различными инструментальными средствами. Используемые при этом подходы различаются поддерживаемым уровнем абстракции (мета уровень, метамета уровень и т.п.) и функциональной направленностью учитываемых метаданных. 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 согласовываться с характером технологий, используемых для создания и применения данной коллекции.

8.2 Технологии баз данных

Достигнутое к настоящему времени состояние развития технологий баз данных можно кратко охарактеризовать следующим образом.

Сформировалась широкая сфера *разнообразных приложений*, в том числе нетрадиционных, что привело к созданию новых классов систем баз данных - пространственно-временных и активных систем, систем реального времени, мультимедийных систем, систем баз данных, поддерживаемых в оперативной памяти (in-Memory Database), мобильных систем.

В первую очередь, в связи с потребностями управления экономикой разработаны методология, модели и инструментарий создания *аналитических приложений* систем баз данных - OLAP, Data Mining, Data Warehousing.

Имеются значительные достижения в области *интеграции технологий* баз данных с Web-технологиями, а также с технологиями текстовых систем.

Ослабляется *доминирующая роль реляционной модели* данных и основанных на ней технологий, разработаны и все шире применяются объектные модели, временные модели и, в качестве шага на пути эволюции реляционных технологий к объектным - объектно-реляционная модель с расширяемой системой типов данных.

Каждодневной реальностью стали *распределенные базы данных*. Отработаны методы распределения данных, архитектурные подходы, реализующие принципы многозвенной архитектуры клиент-сервер, архитектуры промежуточного слоя, мобильные архитектуры.

Предложены архитектуры и модели, обеспечивающие *интеграцию неоднородных ИР* в системах баз данных с различной степенью прозрачности их распределения и неоднородности, а также методы повторного использования ресурсов унаследованных систем.

Усовершенствованы *модели транзакций*, создана теория и технология *потоков работ*. Разработаны спецификации связывания языков программирования для реляционных и объектных систем баз данных. Нашли практическое использование модели *безопасности данных*.

Сформировалась культура разработки крупных приложений, неотъемлемой составной частью которой стало использование инструментальных *средств проектирования и разработки*, поставляемых всеми ведущими поставщиками серверов баз данных, в том числе, и средств, основанных на методологиях объектного анализа и проектирования.

Сложился мощный *конкурентный рынок* программного обеспечения систем баз данных, ряд крупных компаний поставляет реляционные, объектно-реляционные и объектные серверы баз данных. Разработаны и используются на практике эффективные *методы хранения и доступа* для временных и пространственно-временных данных, что весьма важно в связи с появлением крупных баз данных терабайтового объема.

Значительные достижения имеются в области создания индустриальных и международных *стандартов* в области баз данных.

В этой связи нужно упомянуть, прежде всего, деятельность Object Data Management Group (ODMG), направленную на выработку стандартов объектных баз данных [47, 67]. Составляющая основу действующего стандарта ODMG объектная модель является расширением объектной модели OMG. Благодаря этому обеспечивается естественное погружение объектных СУБД, соответствующих стандарту ODMG, в архитектуру CORBA. Активизация объектного направления в технологиях баз данных в значительной мере стимулировалась также включением в указанный стандарт спецификаций связываний для объектных языков программирования C++ и особенно Java, а также значительным прогрессом в развитии общей объектной инфраструктуры, в формировании которой определяющий вклад, несомненно, принадлежит консорциуму Object Management Group (см. п. 8.5).

Заметим, что в настоящее время ODMG завершает работу над новой версией поддерживаемого консорциумом стандарта - ODMG 3.0, называемого теперь "стандартом для хранения объектов", в которой уточнены спецификации объектной модели, введен ряд улучшений в связывание для языка Java, внесены некоторые изменения, позволяющие применять стандарт в системах, использующих отображение объектной среды в реляционную [48].

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 предысторией, 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) позволяет факультативно использовать совместно с ядром системы специальные модули DataBlade, позволяющие вводить допол-

нительные типы данных. Один из таких модулей, поставляемых Informix, поддерживает временной ряд как дополнительный тип данных [29] 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* [20]. 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 также основные объектные возможности (структурные типы, определяемые пользователем, средства описания поведения, наследование свойств, уникальная идентификация объектов и др.). Предполагается включить в SQL:1999 спецификации управления данными, внешними по отношению к базе данных (SQL/Management of External Data, SQL/MED) [13]. Будет включена также поддержка темпоральных свойств данных. Конструктивная основа для соответствующих спецификаций (концепции времени, глоссарий временных баз данных, язык TSQL2) была создана еще около пяти лет назад [31, 54], и ведется работа над соответствующим разделом стандарта. В SQL:1999 включаются также спецификации связывания для языка Java (OLB/SQL, см. ниже). Многие элементы нового стандарта уже реализованы в коммерческих серверах баз данных.

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 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. Он определяет спецификации интерфейса прикладного программирования на языке Java для доступа к SQL-базам данных с поддержкой возможностей динамического связывания. Первоначальная его версия входила в состав JDK 1.1.

Создание дополнительных возможностей более высокого уровня для взаимодействия языка Java с системами реляционных баз данных было целью образованного в апреле 1997 г. консорциума (группа SQLJ) в составе компаний Compaq, IBM, Informix, Micro Focus, Microsoft, Oracle, Sun и Sybase. Его участники разработали спецификации SQLJ и предложили их для принятия в качестве официального стандарта ANSI. SQLJ Часть 0 [21] обеспечивает встраивание статических операторов SQL в Java-программы и тем самым доступ из них к базам данных. Эта спецификация вошла как одна из частей в стандарт SQL:1999 под названием "Связывания для объектных языков (SQL/OLB)" и в декабре 1998 г. была одобрена ANSI. В сентябре 1999 г. аккредитованный при ANSI Национальный комитет по стандартам информационных технологий (NCITS) одобрил SQLJ Часть 1 [22] - спецификацию, которая дает возможность создавать хранимые Java-программы для SQL-баз данных. Третья часть проекта - SQLJ Часть 2 "Типы данных SQL, использующие язык программирования Java" находится в настоящее время на рассмотрении в ANSI.

На дальнейшее развитие технологий баз данных в направлениях, представляющих интерес для разработок коллекций для электронных библиотек, наиболее существенное влияние, по нашему мнению, могут оказать: (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, основанных на языке XML, и интеграция технологий баз данных с Web-технологиями; (4) стандартизация интерфейсов между системами баз данных и системами программирования на языке Java.

8.3 Текстовые технологии

Радикальные шаги в развитии технологий этого класса были в значительной мере связаны с появлением недорогих запоминающих устройств большого объема, увеличением производительности процессоров компьютеров массового производства, с созданием эффективной и доступной технологии оцифровки текстовых источников, а также с осознанием потребностей эффективной обработки текстов в глобальной среде Internet.

Главными технологическими достижениями в этой области нужно, вероятно, считать совершенствование текстового поиска, разработку методов семантического анализа естественно-языковых текстов, подходов к семантической интеграции текстовых ИР, наконец, создание полнофункциональных систем управления тексто-

выми документами.

Современные текстовые поисковые системы далеко ушли от по своим возможностям от ранних дескрипторных ИПС. Они допускают *полнотекстовый контекстный поиск* с учетом изменяющихся грамматических форм термов контекста, возможности поиска по фонетической близости (по созвучию) слов, поиска в многоязычных документах, поиска по сложным критериям, включающим не только контекст, но и булевские условия, аргументами которых могут быть различные атрибуты текстовых документов, например, время публикации. Последние достижения в области поиска связаны с *семантическим анализом* текстов документов и текстов запросов, построением их поисковых образов (например, в виде семантических сетей), семантическая близость которых оценивается и с применением различных эмпирических метрик. Это позволяет, например, использовать в качестве запроса заданный полнотекстовый документ.

Одна из разновидностей технологий такого рода предложена в работе [74]. Она обеспечивает смысловой поиск и автоматическое индексирование текстовых документов без использования смыслового тезауруса. Поисковым образом текста служит при этом множество его наиболее сильно связанных слов. Этот набор определяется на основе сопоставления текста со множеством заданных для предметной области обучающих текстов. Используемая для оценки силы связей слов метрика основана на комбинаторной статистике словоупотребления в данном тексте.

Функции полнотекстового поиска стали поддерживаться в механизмах серверов баз данных ведущих производителей. Полнотекстовый поиск получил широкое распространение и в среде WWW благодаря созданию мощных *поисковых серверов* и предоставлению к ним свободного доступа. В последнее время уделяется большое внимание повышению целенаправленности поиска и селективных возможностей поисковых механизмов таких серверов. Именно этой цели служат рассмотренные выше (см. п. 8.1) средства *описания семантики* документов в WWW [51, 52, 60, 61], в частности, предоставляемых средствами инфраструктуры нового языка разметки XML.

Успехи в семантическом анализе текстов позволили расширить возможности полнотекстовых систем и обеспечивать их средствами не только функции поиска, но и автоматическое аннотирование документов, их классификацию и кластеризацию. Возникли предпосылки для формирования нового пласта технологий в области семантической обработки текстов, которые часто называют *глубинным анализом текста* (Text Mining) [55], по аналогии с технологиями Data Mining в управлении данными. Общий их смысл состоит в обеспечении для пользователя возможностей восстановления той содержательной информации, которая была заложена в текст документа его создателем. Технологии Text Mining активно развиваются в последнее время. Одним из лидеров в этой области является компания IBM.

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 которых 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 одной из напередзаданных категорий с одновременным его индексированием для полнотекстового 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 для указанных целей.

Другое важное направление в области текстовых технологий связано с созданием полнофункциональных *систем управления документами* [63, 65, 70]. В настоящее время поставляются разнообразные программные продукты этой категории. Наиболее развитые из них основаны на интеграции технологий баз данных, Web-технологий, коммуникационных возможностей, а также технологий Java.

Основные функции развитых систем управления документами - поддержка процесса разработки документов, их хранения, технической и семантической 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 включа-

ют средства 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 аннотирование 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 семантического полнотекстового поиска. В распространении документов используются push- и pull-технологии, электронная почта и другие средства.

Интенсивное развитие коммуникационной среды открыло возможности для интеграции распределенных документальных ИР в рамках единых коллекций. Имеется ряд подходов к решению этой задачи, обеспечивающих различную степень семантической интеграции ресурсов, в зависимости от того, какой уровень интероперабельности обеспечивает описание их семантики. Один из таких подходов основан на *интеграции тезаурусов* интегрируемых коллекций. Известны некоторые экспериментальные проекты в этой области. Например, в [33] предлагается подход к решению этой задачи на основе использования медиаторов в среде архитектуры CORBA. Другие подходы к семантической интеграции ИР основаны на использовании онтологических контекстов ресурсов и на их интеграции [69].

Следует здесь упомянуть также о новых возможностях доступа пользователей к документальным коллекциям, обеспечивающего функции информационного поиска на основе *стандарта Z39.50* [39]. 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 [27]. Кроме того, предусматривается возможность поиска по элементам Дублинского ядра [35].

Если попытаться сформулировать важнейшие факторы, определяющие дальнейшее развитие текстовых технологий, то, на наш взгляд, следует отнести к их числу появление развитой коммуникационной среды, интеграцию текстовых технологий в среду Web, которая в силу огромных объемов доступных в ней ресурсов и гигантского количества пользователей предъявляет высокие требования к селективным возможностям поисковых серверов, а следовательно, требует высокого уровня семантической поддержки. Не случайно, как отмечалось в п. 8.1, в настоящее время интенсивно ведутся работы по созданию средств семантического описания ресурсов WWW. Важным стимулирующим фактором развития текстовых технологий становится также их внедрение в сферу управления корпорациями, что привело к созданию полнофункциональных систем управления документами.

8.4 Web-технологии

Воплощением достижений Web-технологий является самая крупная из существующих информационных систем - *World Wide Web*. Ее характеристики в полной мере характеризуют состояние развития этого пласта технологий.

WWW представляет собой глобальную открытую бесконечно масштабируемую распределенную гипермедийную систему, распределение и неоднородность ресурсов которой прозрачны для пользователей. Система обладает огромными интенсивно наращиваемыми информационными ресурсами, большинство из которых предоставляется для свободного доступа в любой момент времени. Среда системы способна интегрировать другие важные группы технологий - технологии баз данных, текстовые технологии, объектные технологии (например, OMG CORBA [14]), технологии Java.

Благодаря всем этим возможностям среда WWW стала *эффективной платформой* для реализации новых важных классов приложений, таких как электронные библиотеки, системы электронного бизнеса и др. Технологии Web нашли также широкое применение в корпоративных информационных системах (интранет).

Наиболее существенное влияние на дальнейшее развитие Web-технологий оказывают, на наш взгляд, принятый консорциумом W3C *стандарт XML* (Extensible Markup Language) [23] - спецификации нового расширяемого языка разметки, а также создание его инфраструктуры - ряда дополняющих XML языковых спецификаций, позволяющих задавать метаданные, описывающие структуру, содержание и другие свойства XML-документов.

Язык XML представляет собой удобное для реализации подмножество известного языка SGML (стандарт ISO). XML - это *метаязык*. На его основе можно конструировать разнообразные языки разметки для различных сфер при-

менения, и эти языки имеют статус приложений XML. Одну из таких конкретизаций XML представляет язык HTML. Именно это обстоятельство обеспечивает возможности для дальнейшего использования в новой среде накопленных огромных HTML-ресурсов.

На основе языка XML уже созданы различные инструментальные средства в статусе его приложений, прежде всего, языки разметки, ориентированные на конкретные предметные области. Можно упомянуть, в частности, научные приложения - язык разметки математических текстов MathML (Mathematical Markup Language) [41], химический язык разметки CML (Chemical Markup Language) [44, 45] и написанный на Java браузер, позволяющий работать с документами на CML, язык разметки для описания астрономических инструментов Astronomical Instrument Markup Language (AIML) [15]. Кроме того, как отмечалось выше, средствами XML DTD разработано определение типов XML-документов для коллекций исторических документов [3].

Интеграция технологий баз данных с Web-технологиями оказалась взаимно востребованной и представляет собой важный фактор, оказывающий влияние на развитие обеих этих групп технологий и стимулирующий их встречное движение. Появляются новые версии СУБД, которые обладают дополнительными функциональными возможностями, ориентированными на использование этих систем в среде Internet. Примером могут служить СУБД Oracle8i или Informix Internet Foundation 2000, уже поставляемые компаниями-разработчиками.

Со стороны Web теперь уже речь идет не о широко использовавшемся "механическом" встраивании систем баз данных в среду WWW с доступом через HTML-формы и CGI. Найдены более фундаментальные подходы, предусматривающие *концептуальное сближение* рассматриваемых технологий благодаря реализации в среде Web основных принципов управления данными, традиционных для систем баз данных, и созданию систем управления слабоструктурированными данными, таких как информационный сервер Tamino компании Software AG.

При этом имеется в виду разработка адекватных моделей данных для Web и более общих *интегрирующих моделей*, позволяющих одновременно оперировать как базами данных, так и слабоструктурированными данными Web. Предусматривается явная поддержка метаданных в Web для различных приложений, разработка языков запросов для Web, а также архитектурных подходов, обеспечивающих интеграцию Web с системами баз данных и основанных, в частности, на принципах архитектуры промежуточного слоя, на 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 [24, 40]. В [50] обсуждается разработанная в Стенфордском университете объектная модель данных Object Exchange Model (OEM), обеспечивающая представление как структур данных в базах данных, так и слабоструктурированных данных Web. В работе [4] представлена разработанная для тех же целей в рамках проекта ARANEUS модель ADM. Рекомендации W3C [17] 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 выше (п. 8.1), 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 [7], WebOQL [6] и целый ряд других [24]. Несколько проектов языков запросов - претендентов на стандартизацию рассматривается также в W3C [16, **?**]. Последний из них (XQL) реализован, в частности, в информационном сервере Software AG Tamino.

Предпринимаются шаги, направленные на *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 (IIOP), являющиеся частью стандарта CORBA [14]. Наконец, OMG принят стандарт основанного на XML языка XMI [57] для обмена метаданными между средствами объектного анализа и проектирования (см. п. 8.1). Некоторые средства для рассматриваемых целей предусматриваются также в CORBA 3 - новой версии стандарта OMG CORBA (см. п. 8.5).

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

8.5 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 поддержки таких сред, базирующаяся на объектной модели OMG и языке определения интефейсов IDL [14, 68], который независим от языков программирования, используемых для реализации приложений. Стандартизованы отображения спецификаций IDL в языки Ada, C, C++, COBOL, Java, Smalltalk. Создан безопасный протокол IIOP 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 [14] - принята OMG летом 1999 г.

В настоящее время OMG завершает разработку *стандарта CORBA 3* [53], все составные части которого уже существуют. 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 межсетевого фильтра (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 со статическим и динамическим вариантом для каждого из них. 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. В новом стандарте предусматривается три специальных конфигурации CORBA - минимальная, предназначенная, главным образом, для встроенных систем, конфигурация для систем реального времени и помехоустойчивая конфигурация.

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 по стандартизации в области *объектного анализа и проектирования*, касающаяся прежде всего средств моделирования метаданных и обмена метаданными между различными CASE-системами (см. п. 8.1). Широкое признание получил стандартизованный OMG визуальный *язык моделирования UML* (Unified Modeling Language) [56], 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 [66, 73, 75]), 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 (исследования геномов и структурная биология), химической информатике, медицинских клинических испытаниях, вычислительной химии [37].

Предполагается прежде всего сформировать основанную на технологии CORBA единую общую архитектуру приложений для проведения исследований в указанных областях [18] и для ее реализации разработать комплекс стандартов OMG (см. например [8, 10, 12, 25]. 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 [38]. 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 изда-

ния, 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.

9 Заключение

Предпринятая в данной работе попытка систематического анализа свойств коллекций информационных ресурсов в электронных библиотеках представляется весьма важной. Автор надеется, что предложенная здесь систематика коллекций позволит их разработчикам получить более четкое и полное представление о требованиях к ожидаемым результатам, возникающих проблемах, возможных подходах, а также о существующих стандартах и технологиях.

Список литературы

[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] *ACM SIGMOD Anthology.* [http://www.acm.org/sigmod/dblp/db/anthology.html]

[3] *American Memory DTD for Historical Documents.* [http://lcweb2.loc.gov/ammem/amdtd.html]

[4] Anzeni P., Mecca G., Merialdo P. *Semistructured and Structured Data in the Web: Going Back and Forth.* SIGMOD Record, V. 26, No. 4, December 1997.

[5] Arms C.R. *Historical Collections for the National Digital Library.* D-Lib Magazine, April 1996.

[6] Arocena G., Mendelson A. *Viewing Web Information Systems as Database Applications.* Comm. of the ACM, July 1998.

[7] 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/wsq-l/PAPER267.html]

[8] *Biomolecular Sequence Analysis.* RFP Response. Initial Submission. OMG Document lifesci/98-10-04.

[9] *CASE Data Interchange Format (CDIF) - Overview.* Electronic Industries Association. CDIF Technical Committee. January 1994.

[10] *Chemical Entity Representation and Interface Definition. Request for Information.* OMG Document lifesci/99-03-05.

[11] Chen P.P. *The entity-relationship model – toward a unified view of data.* ACM TODS, 1(1): 9-36, March 1976.

[12] *Clinical Trials Workgroup White Paper, Draft B.* OMG Document lifesci/98-06-01.

[13] Committee Draft *Database Language SQL - Part 9: SQL/MED.* December 1998. [ftp://jerry.ece.umassd.edu/isowg3/dbl/YGJdocs/ygj023.pdf]

[14] *Common Object Request Broker Architecture.* Version 2.3. Object Management Group, June 1999. OMG Documents formal/99-07-01 – formal/99-07-28.

[15] Cover R. *Astronomical Instrument Markup Language (AIML)* [http://www.oasis-open.org/cover/aiml.html]

[16] Deutsch A., Fernandez M., Florescu D., Levy A., Suciu 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]

[17] *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]

[18] *Domain Software Architecture for Life Sciences Research.* OMG Document lifesci/99-03-07.

[19] *Dublin Core Metadata Element Set Reference Description*, Version 1.1, 1999-07-02. [http://purl.org/dc/documents/proposed_recommendations/pr-dces-19990702.htm]

[20] 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.

[21] 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.

[22] Eisenberg A., Melton J. *SQLJ - Part 1: SQL Routines using the Java Programming Language*. SIGMOD Record, Vol. 28, No. 4, December 1999.

[23] *Extensible Markup Language (XML) 1.0*. W3C Recommendation 10-February-1998. [http://www.w3.org/TR/1998/REC-xml-19980210]

[24] Florescu D., Levy A., Mendelzon 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.

[25] *Genomic Maps RFP*. OMG Document lifesci/98-11-07.

[26] Griffin S.M. *NSF/DARPA/NASA Digital Libraries Initiative*. D-Lib Magazine, July/August 1998.

[27] Hammer S., Favaro J. *Z39.50 and World Wide Web*. D-Lib Magazine, March 1996.

[28] *HTML 4.0 Specification*. W3C Recommendation, Revised on 24-Apr-1998. [http://www.w3.org/TR/REC-html40]

[29] *Informix TimeSeries DataBlade Module. User's Guide*. Version 3.1. Informix Software Inc. April 1998.

[30] *ISO/IEC 10027:1990 Information Resource Dictionary System (IRDS) Framework*.

[31] *Jensen C.S., ed. A consensus Glossary of Temporal Database Concepts*. SIGMOD Record, Vol. 23, No. 1, March 1994.

[32] 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.

[33] Kramer R., Nikolai R., Habeck C. *Thesaurus federations: loosely integrated thesauri for document retrieval in networks based on Internet technologies*. International Journal on Digital Library, 1, 1997.

[34] Lagoze C., Fielding D. *Defining Collections in Distributed Digital Libraries*. D-Lib Magazine, November 1998.

[35] 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/~rrl/docs/dublincoreandz3950.html]

[36] Ley M. *Computer Science Bibliography*. Universität Trier. [http://dblp.uni-trier.de/]

[37] *Life Sciences Research Domain Task Force Pocket Guide*. OMG Document lifesci/99-04-02.

[38] *LSR Bibliographic Query Services. Request For Proposal. Draft D*. OMG Document lifesci/99-03-10.

[39] Lynch C.A. *The Z39.50 Information Retrieval Standard*. D-Lib Magazine, April 1997.

[40] Manola F. *Toward a Web Object Model*. Object Services and Consulting, Inc. February 10, 1998. [http://www.objs.com/OSA/wom.htm]

[41] *Mathematical Markup Language (MathML) 1.0 Specification*. World Wide Web Center Recommendation 07-April-1998. [http://www.w3.org/TR/1998/RECMathML-19980407]

[42] *Meta Object Facility (MOF) Specification*. October 7, 1997. Joint Revised Submission. OMG Document ad/97-10-02.

[43] *Metadata Interchange Specification (MDIS)*. Version 1.1. Meta Data Coalition, August 1, 1997.

[44] *Design Consideration for CML*. [http://www.xml-cml.org/design.html]

[45] Murray-Rust P. *JUMBO and XML/CML demonstration*. [http://www.nottingham.ac.uk/~pazpmr/README]

[46] *Namespaces in XML*. World Wide Web Consortium 14-January-1999. REC-xml-names-19990114. [http://www.w3.org/TR/1999/REC-xml-names-19990114/]

[47] *Object Database Standard: ODMG 2.0*. Ed. by R.G.G. Cattell, D.K. Barry. Morgan Kaufmann Publishers, Inc. 1997.

[48] *ODMG 3.0 to be Published*. [http://www.odmg.org/frrbottom.htm]

[49] *Open Information Model (OIS)*, Version 1.0. Review Draft. Meta Data Caolition. April 12, 1999.

[50] Papakonstantinou Y., Garsia-Molina H., and Widom J. *Object Exchange Across Heterogeneous Information Sources*. IEEE Int. Conf. on Data Engineering, Taipei, March 1995.

[51] *Resource Description Framework (RDF). Model and Syntax Specification*. W3C Recommendation 22 February 1999. [http://www.w3.org/TR/REC-rdf-syntax/]

[52] *Resource Description Framework (RDF). Schema Specification*. W3C Proposed Recommendation 03 March 1999. [http://www.w3.org/TR/PR-rdf-schema/]

[53] *Sigel J. What's Comming in CORBA 3*. [http://www.omg.org/news/pr98/component.html]

[54] 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.

[55] *Text Mining Technology. Turning Information Into Knowledge. A White Paper from IBM*. Ed. by D. Tkach. IBM Software Solution. February 17, 1998.

[56] *UML Specification*. OMG Documents ad/97-08-02 – ad/97-08-09.

[57] *XML Metadata Interchange (XMI)*. Proposal to the OMG OA&DTF RFP3: Stream-based Model Interchange Format (SMIF). OMG Document ad/98-10-05.

[58] *XML Schema Part 1: Structures*. W3C Working Draft 5, November 1999. [http://www.w3.org/TR/1999/WD-xmlschema-1-19991105]

[59] *XML Schema Part 2: Datatypes*. W3C Working Draft 5, November 1999. [http://www.w3.org/TR/1999/WD-xmlschema-2-19991105]

[60] Weibel S. *Metadata: The Foundations of Resource Description*. D-Lib Magazine, July 1995.

[61] Weibel S. *The State of the Dublin Core Metadata Initiative*. April 1999. D-Lib Magazine, April 1999.

[62] *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.

[63] Баласанян В. *Концепции системы автоматизации отечественного документооборота*. Открытые системы, 1, 1997.

[64] Булах Е., Кузина И., Храмцов П. *Развитие стека спецификаций W3C или гносеология XML*. Открытые системы, 5-6, 1999.

[65] Гавердовский А. *Концепции построения систем автоматизации документооборота*. Открытые системы, 1, 1997.

[66] Захаров В.Н. *Создание интегрированных электронных библиотек на основе неоднородных распределенных электронных коллекций научной информации*. Институт проблем информатики РАН. Проект РФФИ 98-07-91061.

[67] Калиниченко Л.А. *Стандарт систем управления объектными базами данных ODMG-93: краткий обзор и оценка состояния*. СУБД, 1/1996.

[68] Калиниченко Л.А., Когаловский М.Р. *Стандарты OMG: Язык определения интерфейсов IDL в архитектуре CORBA*. СУБД, 2/1996.

[69] Калиниченко Л.А., Скворцов Н.А., Брюхов Д.О., Кравченко Д.В., Чабан И.А. *Подход к проектированию персонализованных электронных библиотек над Web-сайтами со слабоструктурированными данными*. Программирование, 3, 2000.

[70] Клименко С.В., Крохин И.В., Кущ В.М., Лагутин Ю.Л. *Электронные документы в корпоративных сетях*. - М.: Анкей - Экотрендз, 1999.

[71] Когаловский М.Р. *Абстракции и модели в системах баз данных*. СУБД, 4-5/1998.

[72] Когаловский 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.: ИПИ РАН, 1993.

[73] Колчанов Н.А. *Интегральная электронная библиотека по пространственным структурам и функциям ДНК, РНК и белков (в составе Электронной библиотеки СО РАН)*. ИЦИГ СО РАН. Проект РФФИ 98-07-91078.

[74] Крейнес М.Г. *Смысловой поиск и индексирование текстовой информации в электронных библиотеках: информационная технология "ключи от текста"*. Электронные библиотеки, Том 2, Вып. 3, 1999. [http://www.iis.ru/el-lib/1999/199903]

[75] 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.

[76] Сергиевская А.Л., Колесниченко Е.Г., Лосев С.А. *Проблемы накопления экспериментальной информации в рамках научной информационной системы.* Электронные библиотеки: Перспективные методы и технологии, электронные коллекции. Первая Всероссийская научная конференция. С.-Петербург, октябрь 1999. Изд. С.-Петербургского госуниверситета, 1999.