ПРИНЦИП ДИНАМИЧЕСКОГО ФОРМИРОВАНИЯ ДОКУМЕНТОВ В ИНФОРМАЦИОННЫХ СИСТЕМАХ, НА ПРИМЕРЕ ИНТЕГРИРОВАННОЙ РАСПРЕДЕЛЕННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ (ИРИС) СО РАН

Шокин Ю.И., Федотов А.М., Леонова Ю.В.
Институт вычислительных технологий СО РАН, Россия, Новосибирск, 630090, пр. акад. Лаврентьева, 6
shokin@ict.nsc.ru, fedotov@ict.nsc.ru, juli@pine.ict.nsc.ru

PRINCIPLE OF DYNAMIC CONSTRUCTION OF DOCUMENTS IN INFORMATION SYSTEMS, ON AN EXAMPLE OF THE INTEGRATED DISTRIBUTED INFORMATION SYSTEM (IDIS) SB RAS

Shokin Yu.I., Fedotov A.M., Leonova J.V.
Institute of Computational Technologies, Russia, Novosibirsk, 630090, Lavrentiev avenue, 6
shokin@ict.nsc.ru, fedotov@ict.nsc.ru, juli@pine.ict.nsc.ru

The paper is devoted to discussion of conceptual principles of construction of the integrated distributed information system on an example of the integrated distributed information system (IDIS) SB RAS. The offered principles are grounded on correspondence of its components to open international standards and usage as protocols of interaction of information subsystems and access to information resources alongside with HTTP of the protocol Z39.50. The designed technology gives possibility to operate with changed in time or depending on access conditions with documents, and also to combine different information resources in the conceptually one information environment.

1. ВВЕДЕНИЕ

Доклад посвящен обсуждению концептуальных принципов построения интегрированной распределенной информационной системы на примере интегрированной распределенной информационной системы СО РАН. Предлагаемые принципы основываются на соответствии ее компонент открытым международным стандартам и использовании в качестве протоколов взаимодействия информационных подсистем и доступа к информационным ресурсам наряду с HTTP протокола Z39.50.

Важнейшей частью Информационной среды Сибирского отделения РАН, создаваемой в рамках целевой программы СО РАН "Информационно-телекоммуникационные ресурсы СО РАН", является информационная поддержка научных исследований, проводимых в Отделении, а также создание и развитие собственных информационных ресурсов, управление

этими ресурсами и обеспечение использования информационных ресурсов мирового научного сообщества, представляемых сетью Internet, распространение своих достижений в виде электронных коллекций, атласов и информационных систем, а также в виде электронных публикаций и электронных библиографических ресурсов. Эти проблемы уже обсуждались в [1], а так же на ежегодных рабочих совещаниях СО РАН по электронным публикациям [4-6].

В Отделении накоплена и постоянно собирается уникальная научная информация, как по различным отраслям наук, так и по природному комплексу. В связи с этим наиболее важной работой, связанной с созданием информационных ресурсов Отделения является создание собственных электронных коллекций, аккумулирующих гигантский научный потенциал Отделения. Для решения проблем информационного обеспечения в Отделении создается "Интегрированная распределенная информационная система СО РАН" (**ИРИС**), в которой бы аккумулировалась большая часть необходимой для сотрудников информации, включая полнофункциональную систему об интеллектуальном потенциале Отделения и "Электронную библиотеку Сибирского отделения РАН" [7,8]. ИРИС представляет распределенную информационную систему об институтах, сотрудниках, научных разработках, публикациях, достижения и др. аспектах, связанных с работой Отделения. ИРИС обеспечивает систему работы с документами различного происхождения (объединение распределенных и локальных электронных информационных и программно-алгоритмических ресурсов, включая документооборот), систему электронной поддержки сбора и накопления информации (системы электронных коллекций, баз данных и т.п.). Принципы, заложенные в проектирование системы, позволяют автоматизировать процессы создания электронных коллекций, библиотек и т.п.

Основное назначение ИРИС связано с созданием единой распределенной информационной среды Отделения, объединяющей в интегрированное информационное пространство распределенных и локальных электронных ресурсов (информационных, программных, алгоритмических) организаций Отделения и комплекса программно-технических средств, обеспечивающего использование этих ресурсов и полнофункциональное управление ими. Создание системы связано с информационной поддержкой исследований по фундаментальным и прикладным направлениям, проводимым в институтах Отделения, а также межинститутских междисциплинарных научных исследований.

Организационно-технологическое обеспечение процесса создания полнофункциональной информационной системы включает в себя большой спектр работ, связанных с организацией системы доступа пользователей к информационно-вычислительным ресурсам и к базам данных, сохранение, поддержку и создание информационных ресурсов Отделения, что самое главное воспитание нового пользователя, способного жить и рабо-

тать в современном информационном мире. Из первоочередных задач, которые решаются в настоящий момент, отметим следующие:

- Инвентаризация существующих информационных ресурсов и включение их в ИРИС.
- Адаптация существующих разработок институтов СО РАН в области построения распределенных систем и имеющихся баз данных для ИРИС.
- Разработка корпоративных стандартов хранения, поиска и представления информации на основе существующих международных и отечественных стандартов.
- Разработка и адаптация технологий коллективной работы исследователей.

Создаваемая информационная система базируется на информационном WWW сервере Отделения (http://www.sbras.ru/), который является интегрирующим звеном для всей системы поддержки информационных ресурсов Отделения.

2. ДОКУМЕНТЫ

В основу создания ИРИС СО РАН положена концепция *"динамической модели формирования документов"*. Используемая концепция основана на расширенной объектной модели документа. Специфика применения объектно-ориентированного подхода для организации и управления информационными документами и ресурсами потребовала уточненного толкования классических концепций и некоторого их расширения. Это определяется потребностями долговременного хранения объектов во внешней памяти, ассоциативного доступа к объектам, обеспечения согласованного состояния в условиях множественного доступа и тому подобных возможностей, свойственных базам данных.

В целом, конструируя технологию описания документов, мы основывались на методике RDF, которая предлагается консорциумом W3C в качестве стандарта для определения и обработки метаданных сетевых информационных ресурсов. Специфика RDF состоит в том, что механизмы описания ресурсов не делают никаких предположений относительно специфики предметной области и могут быть удобны для описания и обработки сведений о любой области. Примечательной стороной RDF является то, что он позволяет сделать утверждения не только о документах (ресурсах), но и о самих утверждениях.

В информационном пространстве события, факты и любые другие сущности реального мира существуют только в форме документов. Вследствие этого документ является основным объектом, с которым оперирует любая информационная система. Под **документами** мы понимаем любое

описание реальной сущности (объекта, факта или понятия), которые составляют информационное наполнение системы.

В основе реализации системы лежит метамодель, исходящая из того, что документ характеризуется набором присущих ему атрибутов и методов, характеризующих связи с другими документами. Информация о документах системы, их атрибутах и методах поддерживается *сервером метаданных*, содержащим метаописания системы и метаописания отдельных коллекций. Сервер метаданных является отдельной частью системы, содержащей описание информационной модели предметной области, параметров настройки стандартных функций системы. По информации сервера метаданных осуществляется динамическая генерация схем базы данных системы и ведение служебных баз данных, в которых хранятся данные, обеспечивающие поддержку стандартных функций системы и динамически определяемые отношения между документами.

В информационной системе реальные сущности существуют либо непосредственно в виде документа, который ее представляет: описывает, представляет или моделирует, либо в виде упоминания об этой сущности, которые присутствуют в других документах, т.е. содержат опосредованную информацию об этой сущности (информация об этой сущности или ссылается на информацию об этой сущности).

Таким образом, под **Документом** понимается целостный информационный объект (в том смысле, как это понимается в языках объектно-ориентированного программирования) фиксированного **Класса**, помещенный в информационное пространство ИРИС, который *описывает, представляет, отображает* или *моделирует* некоторую сущность реального мира. **Класс** документа определяет смысловую структуру документа, атрибуты и функции, а так же методы доступа к нему. В Классе задается функциональность: множество Документов, принадлежащих одному **Класс**у, выполняют одинаковые функции. Допускается порождение нового **Класс**а на основе уже существующего **Класс**а – наследование. В этом случае новый **Класс**, называемый **подклассом** существующего **Класс**а наследует все атрибуты и методы существующего **Класс**а. В подклассе, кроме того, могут быть определены дополнительные атрибуты, функции и методы.

Очевидно, что приведенный выше набор **функций** документа не является исчерпывающим и может быть расширен. Поясним смысл основных функций документа:

Документ-Описание – содержит описание реальной сущности, например, описание некоторой организации или конкретной персоны, информация о которых используется в информационной системе, т.е. содержит информацию о некоторой сущности, но при этом сам не является сущностью. Отметим, что описательными документами

также являются описания информационных ресурсов в каталоге или описание библиографических источников.

Документ-Представление – непосредственно является конкретной сущностью, например, научная статья, книга или фотография. Кроме того, к этому классу относятся документы, которые содержат информацию о некоторой сущности и при этом сами является сущностью, например, библиографические карточки.

Документ-Отображение – является отображением другого документа, т.е. его точной копией или электронным образом, например, PDF (или PS)-файл статьи, сканированный документ и т.д.

Документ-Моделирование – моделирует некоторую реальную сущность, например, с помощью компьютерной программы.

Отметим, что документами в нашем понимании так же являются компьютерные программы, алгоритмы, растровые и векторные изображения и их описания. Научная статья и ее библиографическое описание (библиографическая карточка) в нашем понимании являются различными, но связанными документами.

3. СТАТУС ДОКУМЕНТА

Помимо принадлежности к Классу Документ обладает Статусом. Статус определяет состояние документа (статичность, версия и т.п.), возможность создания КОПИЙ документа и/или наличие оригинала, наличие авторского и имущественного прав, и т.п.

Права собственности и копии документа

Каждый Документ, представленный в системе имеет “хозяина”: у документа есть *автор, собственник* и *владелец*.

Владелец документа отвечает за хранение и представление документа пользователями системы. В нашем понимании ИРИС является владельцем всех представленных в ней документов. Документ может не принадлежать системе, т.е. его “владельцем” может быть другая информационная система, а в нашей системе содержится только его описание или ссылка на этот документ.

Автор – документ, представленный в системе, может иметь авторство. Это особенно важно при публикации научных электронных коллекций. Автор несет ответственность за содержание документа.

Собственник документа несет ответственность за содержание документа и имеет право пользоваться и распоряжаться принадлежащим ему документом по своему усмотрению (передавать права соб-

ственности). Собственниками документов могут быть организации или лица, зарегистрированные в системе.

Документ характеризуется наличием оригинала и копий или дубликатов. В распределенной системе Документ может быть представлен в различных местах (а так же иметь нелегальные или платные копии).

Оригинал – “первый” экземпляр документа (экземпляр, принадлежащий собственнику или автору).

Копия – документ, полностью воспроизводящий информацию оригинала и все его внешние признаки или часть их.

Иначе говоря, в информационном пространстве фактически существует множество копий документов. Это, в свою очередь, может приводить к появлению самостоятельных копий, живущих своей жизнью. Но понятие самостоятельной копии противоречит самой сути копии, и в контексте ИРИС мы не рассматриваем такие копии. Доступ к документам производится в соответствии с некоторой политикой распределения. Например, в зависимости от местоположения пользователь получает доступ к локальной или “ближайшей” удаленной копии документа.

С общесистемной точки зрения любой документ уникален, не может существовать идентичных документов, нескольких экземпляров документа. Оригинал и копия документа – это разные документы. Поэтому в системе документ определяется однозначно в соответствии с его статусом. Что касается документов, владельцами которых является система, то система обеспечивает идентичность оригинала и копии, что касается “чужих” документов, система обеспечивает проверку их актуальности.

Копия документа является собственностью владельца документа-оригинала. Копия как документ должна иметь дополнительные атрибуты, позволяющие идентифицировать не только автора, но и владельца-создателя копии и т.д.

Статичность документа и версии документов

Документ в информационном пространстве не является застывшим объектом. Документ может передаваться для обработки другим пользователям, над документом выполняются операции, которые могут менять его состояние или значения его свойств, удалять документы и создавать новые документы. Т.е. у документа есть определенный жизненный цикл. Понятие жизненного цикла документа стало, в определенном смысле, уже стандартом. Жизненный цикл документа включает в себя следующие фазы:

- Создание/ввод документов
- Модификация документов
- Утверждение документов
- Опубликование документов

- Повторное использование документов
- Устаревание и передача документов в архив

Статичность документа отражает подходящее состояние (например, этап жизненного цикла) документа. Например, документ может сначала быть идентифицирован как “Release” (рабочий). После утверждения, состояние может быть установлено в “Current” (принятый) или позже в “Stable” (изданный). Каждому состоянию документа можно сопоставить отдельные права доступа на документ (к примеру, при переходе из состояния “Current” в “Stable” права доступа на редактирование документа будут удалены).

Коллективный характер работы с документами, требование повторного использования содержащихся в них сведений выводят в число базовых характеристик управление версиями документов, т.е. хранение всех промежуточных вариантов с историей модификаций и возможность порождения нового документа на основе любой из существующих версий.

Для документа может существовать множество версий. Каждая версия документа идентифицируется первичным ключом <UIDV> – совокупностью атрибутов, уникально идентифицирующую версию документа. Каждой версии документа присваивается уникальный внутренний идентификатор (UID). Но UID не следует отождествлять с идентификатором документа. Значение идентификатора документа остается постоянным для любой версии документа, в то время как при создании новой версии документа генерируется новое значение UID. Идентификатор документа используется для связывания документа с другими документами. Номера версий генерируются автоматически при создании версий. Номера версиям назначаются в порядке возрастания и никогда не используются повторно. Это отражает семантику версий: версия с большим номером является более новой; отсутствие версии с указанным номером (меньшим максимального) означает, что эта версия была уничтожена. Тип версии идентифицирует состояние версии документа (Release, Current, Stable).

Все версии документа порождаются не только линейно, но и в виде дерева версий, которое получается в случае возврата к какой-либо версии и порождении от нее ответвления версии. Такое порождение может привести к появлению и развитию параллельных версий документа. Примерами таких документов являются алгоритмы или проекты.

Одна из версий документа является текущей, то есть действительной на данный момент. При обращении пользователя к документу рассматривается именно текущая версия. Но всегда можно обратиться к любой конкретной версии документа. Каждая новая версия документа создаётся как текущая версия. Это гарантирует, что пользователь читает или редактирует самую свежую версию документа.

Права доступа к документу

Право доступа разрешает пользователю исполнять определённый набор действий над документом. ИРИС имеет функциональные возможности, чтобы установить дифференцированные права доступа для групп или индивидуальных пользователей к документам. Определенные права доступа для документа могут быть назначены для индивидуальных пользователей или группы пользователей. С другой стороны, документы могут также быть сделаны доступными для анонимного доступа из Internet. Документ всегда связан с определенным пользователем — своим *собственником*. Права доступа приписываются документу. Собственнику разрешено изменять права доступа к документу. Имеется три стандартных набора прав доступа к документам: административный, служебный и публичный.

Публичный тип доступа предоставляется любому пользователю Internet. При этом типе доступа пользователь может просматривать открытые для просмотра документы, осуществлять простой и квалифицированный поиск документов.

Служебный тип доступа предоставляется пользователю при условии обязательной регистрации в системе. Этот тип доступа позволяет пользователю просматривать информацию, закрытую для публичного просмотра, и осуществлять простой и квалифицированный поиск документов.

Административный тип доступа предоставляется пользователю при условии обязательной регистрации в системе. Административный тип доступа позволяет редактировать, удалять и создавать новые документы. Администратор может осуществлять управление версиями, изменять состояние версий и удалять ненужные версии.

Для изменения документ забирается на редактирование и блокируется, а по окончании редактирования возвращается в базу. До возврата в базу другие пользователи не могут редактировать документ.

Если права доступа к документам в системе можно условно разделить на две части: *навигация* (просмотр однотипных документов в виде различного рода списков, выбор объектов) и *модификация* (создание, редактирование, удаление документов), то первый вид прав характеризует простого пользователя и пользователя со служебным доступом, второй – администратора. Унификация доступа к Документам, реализуемая в ИРИС основана на том, что любой документ должен всегда принадлежать некоторому фиксированному КЛАССУ.

4. ОБЪЕКТНАЯ МОДЕЛЬ ДОКУМЕНТА

Исходя из объектной модели представления информации, в основе нашей системы лежат “метаданные” — это структурированные сведения о документе или ресурсе, представляющие его свойства (атрибуты) и функции. На основе метаданных осуществляется поиск документов (ресурсов), вывод результатов поиска, управление ресурсами, взаимодействие с ними. Формальное определение смысловой структуры Документа дается (мета)описанием Класса документа (аналог DTD), в котором каждый тип документов представляется в виде набора объектов со своими характеристиками и атрибутами. В модели RDF документ рассматривается как частично-упорядоченный набор абстрактных объектов (элементов), обладающих свойствами (атрибутами) и имеющих идентификатор. Любой объект при своем создании получает генерируемый системой уникальный идентификатор, который связан с объектом все время его существования и не меняется при изменении состояния объекта.

RDF позволяет определять произвольные объекты в документе. Атрибуты (имена и значения) должны выбираться из словарей, связанных с теми или иными предметными областями. Формально RDF не накладывает никаких ограничений на значения атрибутов объектов, перекладывая создание соответствующих словарей на заинтересованные организации. Основной словарь имен объектов системы создан на основе словарей стандартных схем метаданных GILS и Dublin Core.

Метаописание Класса документов дает структурные свойства объектов, составляющих документ. При этом структура объекта определяется как линейная последовательность атрибутов и/или иерархий атрибутов. Спецификация DTD требует в описании объекта присутствия следующих атрибутов:

name – имя объекта (используемое в ссылках на объект);

title – название объекта;

request – обязательность объекта;

search – возможность включения в поиск и в навигацию;

template_input (output) – шаблоны ввода/вывода;

order – упорядочение объекта.

В связи с требованиями поддержки электронных коллекций в системе список атрибутов был расширен следующими:

type – тип объекта: накладывает ограничения на содержание объекта и определяет способ работы с объектом;

access – уровень доступа к объекту: используется для разграничения доступа к объекту в соответствии с правами пользователей.

Каждый объект имеет состояние, поведение и содержание. *Состояние объекта* – набор значений его атрибутов, значение атрибута объекта – это тоже некоторый объект или множество объектов. *Поведение объекта* –

набор методов доступа (программный код), оперирующих над состоянием и содержанием объекта. *Содержание объекта* – информационное наполнение данного объекта: это может быть ссылкой (link) на объект или на другой документ или на часть другого документа. Состояние и поведение объекта инкапсулированы в объекте; взаимодействие между объектами производится на основе передачи сообщений и выполнения соответствующих методов. Поведение объекта зависит от запроса к документу, т.е. в зависимости от запроса и уровня доступа объект может менять свое содержание.

Типы объектов (элементов) являются спецификациями поведения этих объектов, которые могут модифицироваться для получения нового поведения. Наследование есть механизм модификации поведения, приводящий к эволюции системы. В нашей концепции при добавлении наследования тип превращается в средство, расширяющее существующий тип. Порожденный тип наследует все, связанные с родителем, качества и добавляет к ним свои собственные определяющие характеристики. Без использования иерархии типов для каждого типа пришлось бы задавать все характеристики, которые бы исчерпывающе его определяли. Однако при использовании наследования можно описать тип путем определения того базового множества, к которому он относится, с теми специальными чертами, которые делают тип уникальным. Важная особенность наследования типов состоит в том, что создание производного типа не требует модификации ядра системы.

Производные типы строятся с помощью операций – *конструкторов типов*. Синтаксически конструктор типов представляет собой конструкцию, позволяющую автоматически создавать производный тип объектов из набора базовых типов (строка, число и т.п.) путем указания его структуры и типов компонент. Нами используются следующие виды конструкторов: *перечисление, массив (список), последовательность, запись, объединение* и *функция определения типа.*

Базовой информационной структурой системы является *коллекция.* Коллекция может состоять либо из одного Класса документов, либо быть динамическим (в зависимости от запроса) объединением (join) классов документов. Система предоставляет возможность оперировать с изменяемыми во времени или в зависимости от условий доступа документами. В зависимости от метаописания информационное наполнение коллекции и документа меняется в зависимости от метода доступа к документу.

5. ЗАКЛЮЧЕНИЕ

В основу системы хранения данных положен принцип информационных хранилищ, с учетом поддержки уже функционирующих технологий, например, с использованием этого принципа уже осуществляется интегра-

ция кадровых баз данных институтов Отделения в подсистему ИРИС по научным организациям и сотрудникам Отделения и система ведения электронных коллекций по проблемам биоразнообразия [2, 3].

Разработанная технология предоставляет возможность объединить различные информационные ресурсы в концептуально единую информационную среду, а также оперативно управлять и актуализировать информацию, хранящуюся в разнородных и распределенных по сети базах данных, организовать гибкий поиск, что самое главное — создать достаточно удобный интерфейс для ее наполнения и сопровождения.

Литература

[1] Шокин Ю.И., Федотов А.М. Информационная система Сибирского Отделения РАН // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Второя Всероссийская научная конференция, Протвино, 26-28 сентября 2000 г.: Сб. докл., Протвино, ГНЦ ИФВЭ, 2000, 6-15,
[http://www.protvino.ru/dl2000/reports/pdf/028.pdf]

[2] Байков К.С., Коропачинский И.Ю., Шокин Ю.И., Шумный В.К., Ермаков Н.Б., Колчанов Н.А., Федотов А.М. Электронные коллекции и проблемы биоразнообразия // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Второя Всероссийская научная конференция, Протвино, 26-28 сентября 2000 г.: Сб. докл., Протвино, ГНЦ ИФВЭ, 2000, 58-65,
[http://www.protvino.ru/dl2000/reports/pdf/40.pdf]

[3] Коропачинский И.Ю., Шокин Ю.И., Шумный В.К., Ермаков Н.Б., Колчанов Н.А., Федотов А.М. Электронный атлас "Биоразнообразие животного и растительного мира Сибири".
[http://www-sbras.nsc.ru/win/elbib/bio/].

[4] Материалы IV рабочего совещания по электронным публикациям El-PUB'99. [http://www-sbras.nsc.ru/ws/elpub99/].

[5] Материалы V рабочего совещания по электронным публикациям El-PUB'2000. [http://www-sbras.nsc.ru/ws/el-pub-2000/].

[6] Материалы VI рабочего совещания по электронным публикациям El-PUB'2001. [http://www-sbras.nsc.ru/ws/elpub2001/].

[7] Федотов А.М., Шокин Ю.И. Электронная библиотека Сибирского отделения РАН. //Информационное общество, N2, 2000.

[8] Шокин Ю.И., Федотов А.М. Библиотека, работающая круглосуточно // ЭКО, N6, 2000.