МАТЕМАТИКА

УДК 681.3, 004.91, 004.415.2

А. М. Гудов, С. Ю. Завозкин, А. С. Меньшиков

МОДУЛЬ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ МЕТАДАННЫХ ДОКУМЕНТА В СИСТЕМЕ ЭЛЕКТРОННОГО ДОКУМЕНТООБОРОТА ВУЗА

Для вузов давно актуален вопрос автоматизации хранения, поиска и обработки больших массивов информации, а также обеспечения безопасности её хранения, передачи и возможности совместного использования, что особенно важно, учитывая территориально-распределённую структуру современного вуза со своими филиалами [1].

Помимо автоматизации, в последнее время всё более актуальной становится ещё одна ключевая функция СЭД – интеграция с другими системами, что в первую очередь позволит объединить все подсистемы вуза в одну информационную систему, а в дальнейшем станет основой для реализации более глобальной цели СЭД – создания единой распределенной информационной системы, позволяющей эффективно использовать разнообразные информационные ресурсы и коллекции электронных документов, доступные в удобном для конечного пользователя в виде передачи данных через сеть.

Ключевым объектом СЭД является документ. Под документом мы будем понимать описание реального документа, которое составляет информационное наполнение нашей системы. Документ состоит из двух частей – содержимое и его метаданные (описание).

Метаданные – это описание документа, которое можно передавать другим системам, отражающее текущие характеристики документа, необходимые для работы механизмов, осуществляющих автоматизацию документооборота.

Роль метаданных системы прежде всего состоит:

- в предоставлении возможностей эффективного обнаружения необходимых данных;
- в обеспечении гибких и разнообразных механизмов отбора документов в соответствии с предъявленными требованиями (поисковым запросом);
- в управлении жизненным циклом документов;
- в ускорении процесса доступа к информации за счёт распределённого хранения метаданных и содержимого документа.

Импорт и экспорт документов существенно облегчается, если метаданные системы соответствуют каким-либо из принятых стандартов. Система описания документа в СЭД построена на основе *спецификации IMS* [4] и стандарта метаданных *Дублинского ядра* [5] с учётом информации, полученной при изучении ГОСТ Р 51141-98 «Делопроизводство и архивное дело» [6] и ГОСТ Р 6.30-2003 «Требования к оформлению документов"» [7].

Информация о документе объединена в следующие основные блоки:

- *General* – содержит общую информацию о документе;
- *Classification* – содержит описание документа в соответствии с разработанной классификацией;
- *Lifecycle* – содержит информацию, отражающую развитие и текущее состояние документа;
- *Technical* - содержит информацию, отражающую технические возможности;
- *Relation* - содержит информацию, описывающую взаимодействие с ресурсами. Метаданные документа могут быть как обязательными, так и необязательными. Набор обязательных метаданных для каждого документа определяется его «наименованием».

По направлению все документы СЭД делятся на входящие, исходящие и внутренние. Для импорта входящих документов предусмотрен соответствующий механизм. На функциональной диаграмме первого уровня, представленной на рис. 1, построенной в стандарте IDEF0, за получение входящих документов отвечает блок «Импорт документа», а за отправку исходящих документов отвечает блок «Экспорт документа». Документы могут быть импортированы в СЭД несколькими способами: ручное занесение, импорт из XML файла и автоматическое определение метаданных (рис. 2).

При поступлении содержимого документа и метаданных в виде XML файла система разбора метаданных выделяет из XML документа метаданные и передаёт их вместе с содержимым в блок создания документов.

Остановимся подробнее на способе автоматического определения метаданных. Актуальность механизма автоматического выделения метаданных из электронного документа определяется необходимостью занесения в СЭД большого количества документов, приходящих извне. Документы могут приходить как в виде набора содержимого и метаданных, понимаемых СЭД, так и просто в виде электронного документа произвольного типа. В случае прихода извне электронного документа без метаданных компонент автоматического выделения метаданных из СЭД позволит сэкономить значительное количество времени, а также избежать опечаток, неизбежных при ручном занесении метаданных.

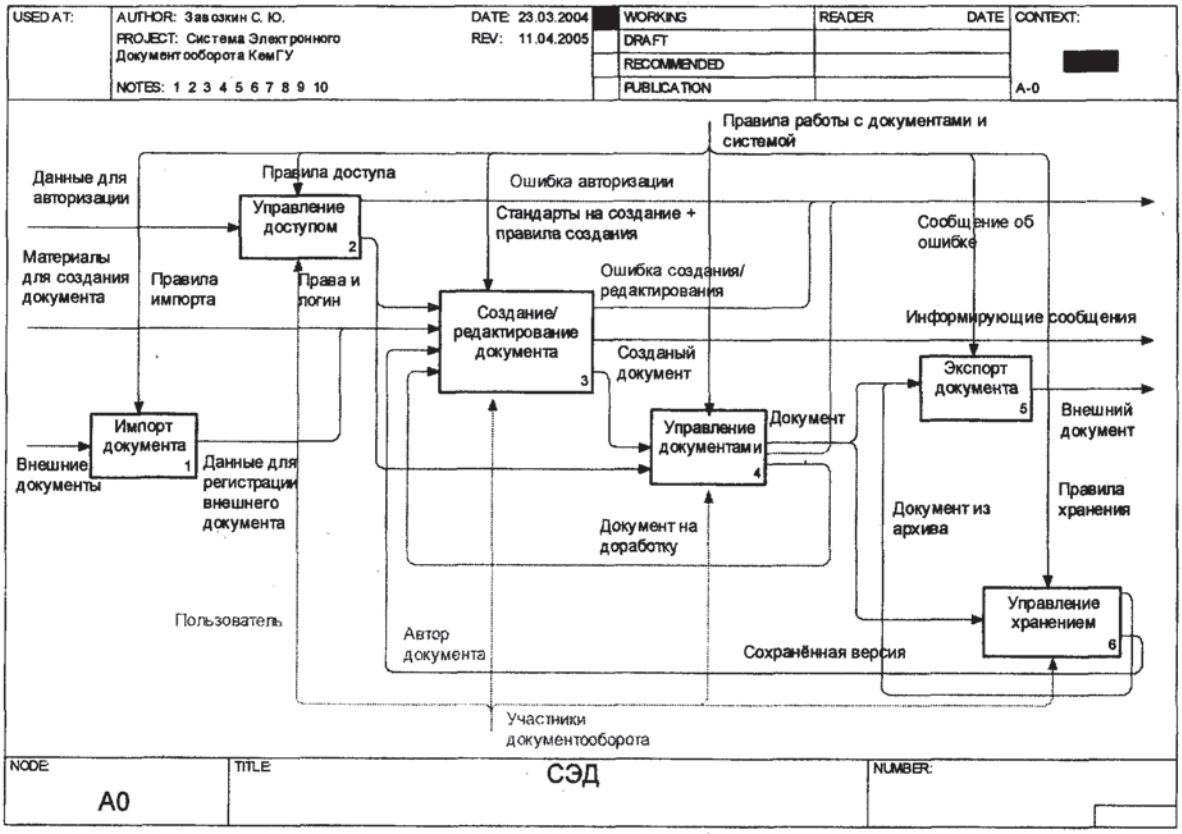

Рис. 1. ***Функциональная модель СЭД***

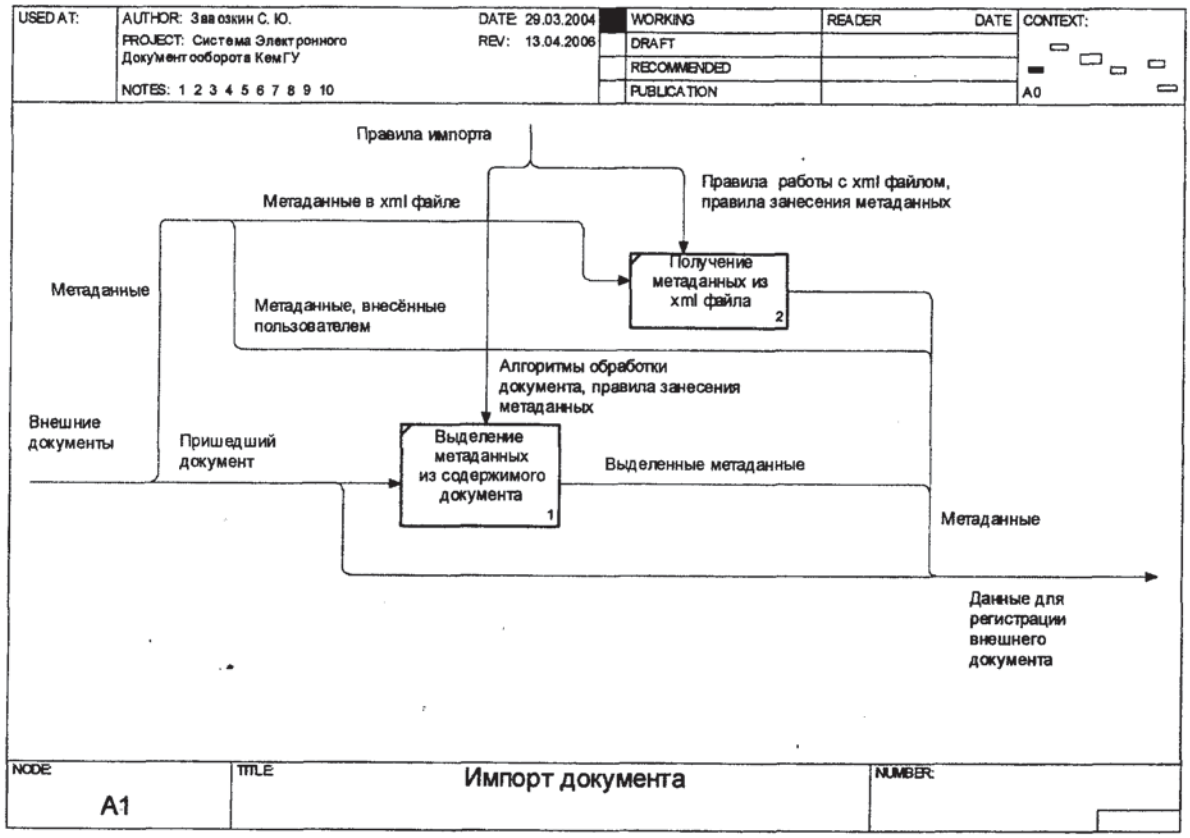

Рис. 2. ***Импорт документа***

Условно задачу автоматического определения метаданных из электронного документа можно разбить на несколько этапов (рис. 3):

1. Определение наименования документа:
   – поиск ключевых слов (например, «приказ», «распоряжение») в тексте документа;

– обработка результатов поиска. Основные полученные данные: количество найденных слов, местоположение найденных слов в тексте (количество символов от начала документа до первого символа найденного слова).

2. Получение списка необходимых для заполнения метаданных, соответствующих выбранному наименованию. Метаданные хранятся в словаре метаданных.
3. Определение самих метаданных. Определение метаданных происходит по тому же принципу, что и определение наименования, т. е. осуществляется:
   – поиск ключевых слов в тексте документа по каждому метаданному из списка;
   – обработка результатов поиска.
4. Представление пользователю результата работы компонента с целью проверки и осуществления правки в случае необходимости.
5. Передача полученных метаданных подсистеме СЭД, формирующей на их основе документ.

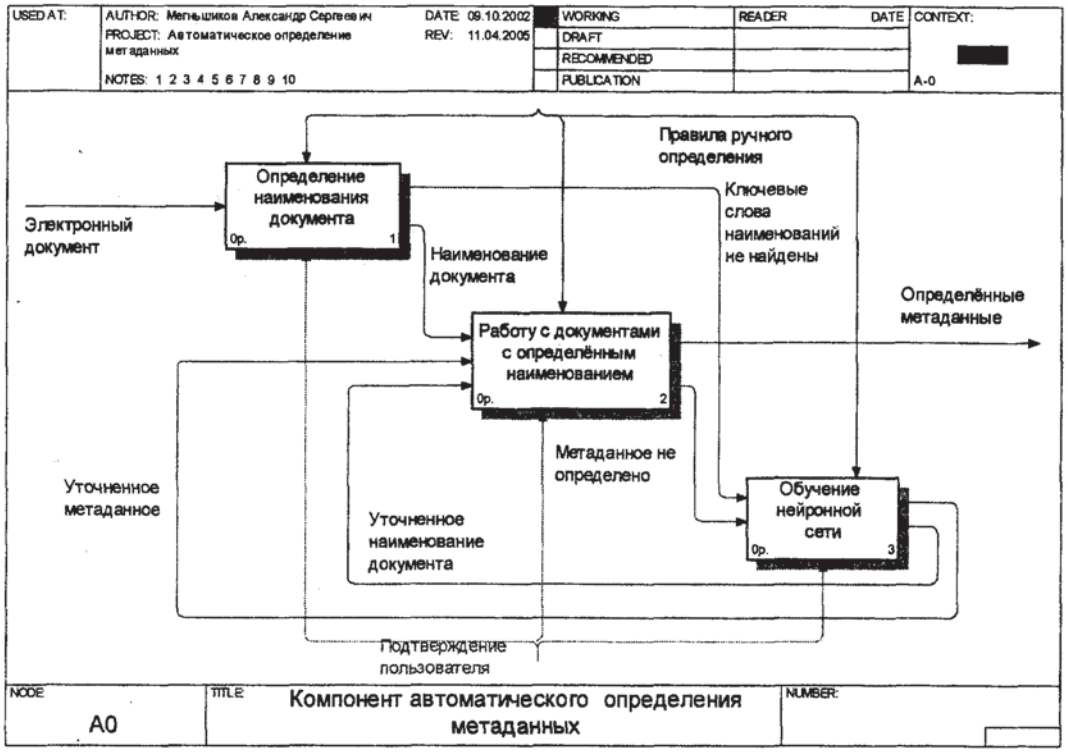

*Рис. 3. **Компонент автоматического определения метаданных***

На основе анализа предметной области был выбран аппарат нейронных сетей как наиболее универсальный, позволяющий решить задачу разработки модели компонента СЭД, осуществляющего автоматическое определение метаданных документа. Рассмотрим его более детально с целью использования для решения нашей задачи.

Искусственные нейронные сети (НС) – совокупность моделей биологических нейронных сетей, представляющих собой сеть элементов – искусственных нейронов, связанных между собой синаптическими (входными и выходными) соединениями [8]. Работа сети состоит в преобразовании входных сигналов во времени, в результате чего меняется внутреннее состояние сети и формируются выходные воздействия.

Основные проблемы, решаемые при помощи аппарата нейронных сетей: классификация образов, кластеризация/категоризация, аппроксимация функций, предсказание/прогноз, оптимизация, память, адресуемая по содержанию, управление. Наша задача относится к задачам классификации образов.

Задача классификации заключается в разбиении объектов на классы, когда основой разбиения служит вектор параметров объекта [9]. Классы зависят от предъявляемых объектов, и поэтому добавление нового объекта требует корректирования системы классов. Будем характеризовать объекты, подлежащие классификации, вектором параметров $x^p \in X$, имеющим N компонент, компоненты обозначаем нижним индексом: $x^p = (x_1^p ... x_N^p)$. Вектор параметров – единственная характеристика объектов при их классификации.

Введем множество классов $C^1..C^M = \{C^m\}$ в пространстве классов C: $(C_1 \cup C_2 \cup .. \cup C_M) \subset C$.

Пространство классов может не совпадать с пространством объектов X и иметь другую размерность. В простейшем случае, когда пространства классов и объектов совпадают, классы представляют собой области пространства X, и объект x^p бу-

дет отнесен к одному из классов m_0, если $x^p \in C^{m_0}$.

Определим ядра классов $\{c^m\} = c^1,...,c^m$ в пространстве классов С, как объекты, типичные для своего класса. Очевидно, что близость объекта к ядру необходимо оценивать численно. Введем меру близости $d(x^p, c^m)$ – скалярную функцию от объекта и ядра класса, которая тем меньше, чем больше объект похож на ядро класса. Задавшись числом классов М, можно поставить задачу классификации: найти М ядер классов $\{c^m\}$ и разбить объекты $\{x^p\}$ на классы $\{C^m\}$, т.е. построить функцию *m*(p) таким образом, чтобы минимизировать сумму мер близости:

$$\min\left\{D = \sum_p d(x^p, c^{m(p)})\right\}.$$

Функция m(p), определяющая номер класса по индексу p множества объектов $\{x^p\}$, задает разбиение на классы и является решением задачи классификации в общем случае [9] (рис. 4).

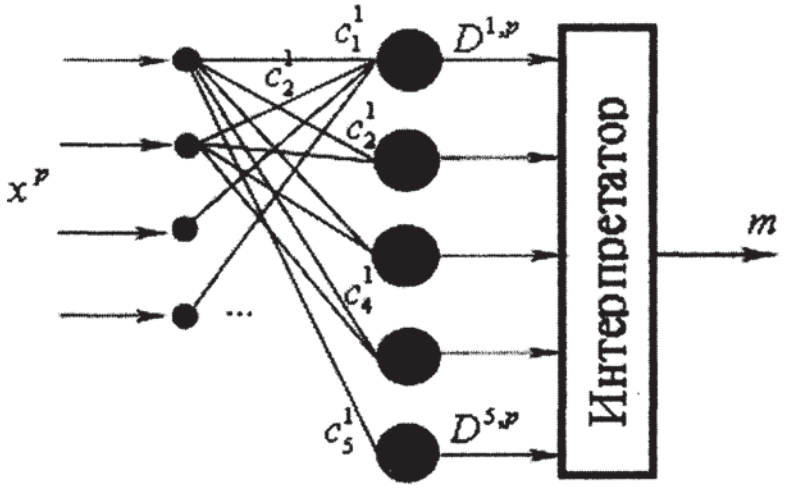

Рис. 4. ***Схема нейронной сети***

Рассмотрим частный случай задачи классификации: предположим, что у нас есть некоторый текст. Текст разбивается на слова и представляется в виде вектора входных параметров.

$x^p = (x_1, x_2, ..., x_n)$, где x_i - слово в тексте, расположенное на i-м порядковом месте в тексте.

Для каждого метаданного определим пространство классов. Например, пространство классов для наименования будет являться C_1 - приказ, C_2 - распоряжение, ... , C_n - новое наименование. В пространстве классов определим ядра для каждого из классов (т.е. наиболее типичных представителей данного класса). Например, для класса «приказ» ядром будет являться

$c_1 = \{"приказ", 10, "слово_1", "слово_2", ...,$
$"фраза_1", "фраза_2", ...\}.$

Здесь слово «приказ», расположенное на 10-м месте в тексте, «слово» – различные слова, часто встречающиеся в документах данного класса, «фраза» - различные фразы, которые часто встречаются в документах данного класса. Таким образом, для каждого класса задано ядро класса.

$c_j = \{"слово", r, (набор\_слов), (набор\_фраз)\}$,

где «слово» соответствует ключевому слову наименования, а r – позиция слова в тексте, (*набор_слов*) и (*набор_фраз*) – это набор часто встречающихся слов и фраз соответственно.

Определим расстояние от x^p до ядра класса c^m : $d(x^p, c^m) = d(x, y, z, k)$, где

$x = \sum_i [NOT("кл.слово" = x_i)]$ – величина, равная 0, если ключевое слово присутствует в документе; $y = |i - r|$ – разность между номерами позиций ключевого слова в документе и ключевого слова из ядра класса, если ключевое слово не найдено, то данная величина равна 0; $z = \sum_i ["слово" = x_i]$ – количество слов документа, совпавших со словами из ядра класса.

$$k = \sum_i \left["фраза_i" = \sum_j x_j \right], j = l, ..., n - m,$$

где l берется сначала 1, затем 2 и т.д. до n-m(n – размерность вектора x, m – количество слов в фразе), данная компонента представляет величину, характеризующую совпадение фраз с фразами ядра класса.

$[\ ]$ – означает переход от логических величин к численным, по правилу: если выражение в скобках принимает истинное значение, то значение всего выражения равно 1, в противном случае все выражение будет равно 0.

После этих преобразований задача определения наименования будет сведена к поиску ми-

нимального расстояния от экземпляра документа до ядра класса $\min(d(x^p, c_i))$.

Выбирается класс, который составляет минимальное расстояние до классифицируемого объекта. Подобным образом будут определены остальные метаданные документа.

В процессе работы нейронная сеть обрабатывает не только стандартизованные документы, но и различные их вариации. Т. е. при изменении шаблона документа нейронная сеть способна обучиться обрабатывать измененные документы.

Процесс обучения рассматривается как настройка архитектуры сети и весов связей для эффективного выполнения специальной задачи. Обычно нейронная сеть должна настроить веса связей по имеющейся обучающей выборке. Функционирование сети улучшается по мере итеративной настройки весовых коэффициентов. Для конструирования процесса обучения прежде всего необходимо иметь модель внешней среды, в которой функционирует нейронная сеть, – знать доступную для сети информацию. Эта модель определяет парадигму обучения [8]. Далее необходимо понять, как модифицировать весовые параметры сети – какие правила обучения управляют процессом настройки. Алгоритм обучения означает процедуру, в которой используются правила обучения для настройки весов.

Существуют три парадигмы обучения: «с учителем», «без учителя» (самообучение) и «смешанная» [8]. В первом случае нейронная сеть располагает правильными ответами (выходами сети) на каждый входной вектор параметров объектов. Веса настраиваются так, чтобы сеть производила ответы как можно более близкие к известным классам. Усиленный вариант обучения с учителем предполагает, что известна только критическая оценка правильности выхода нейронной сети, но не сами правильные значения выхода. Обучение без учителя не требует знания правильных ответов на каждый пример обучающей выборки. В этом случае раскрывается внутренняя структура данных или корреляции между классами, что позволяет распределить классы по категориям. При смешанном обучении часть весов определяется посредством обучения с учителем, в то время как остальная получается с помощью самообучения.

В нашем случае тип обучения нейронной сети – смешанный, обучение предполагает как самообучение нейронной сети посредством статистической обработки и самостоятельной настройки весовых коэффициентов (ядер классов), так и возможность определения нового класса для любого метаданного пользователем. Например, на рис. 5 показан механизм ручного определения нового наименования документа.

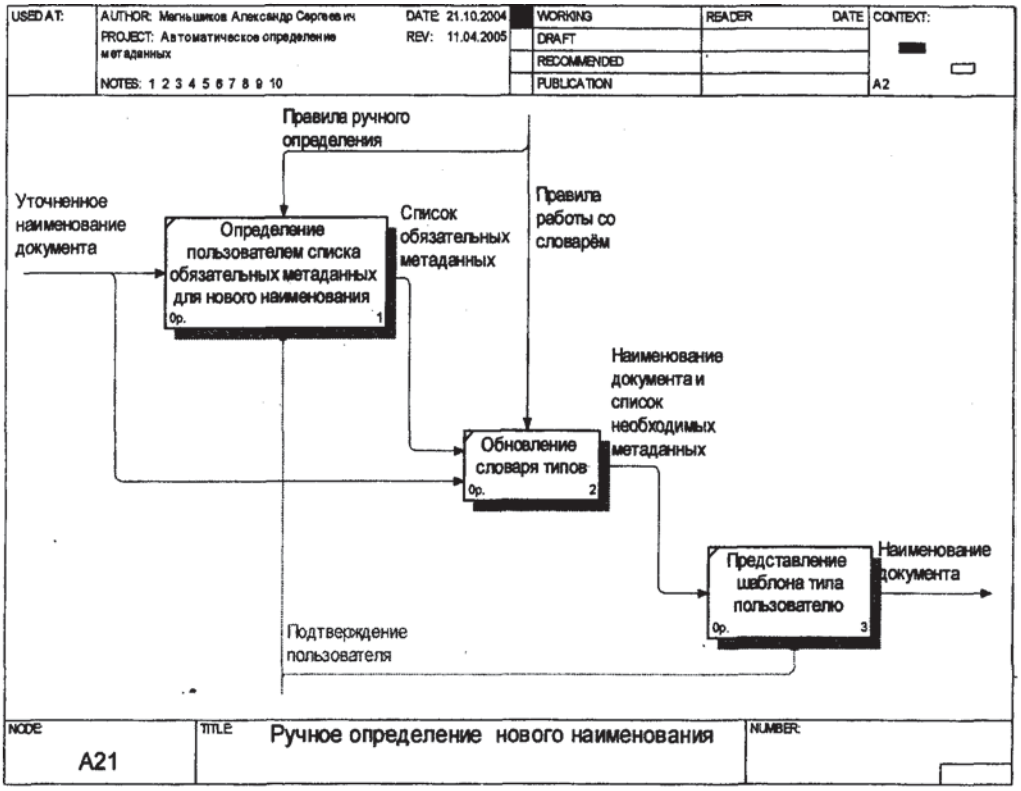

Рис. 5. ***Ручное определение нового наименования***

Определённые метаданные документа будут храниться в реляционной базе данных. Построена модель «сущность – связь», представляющая древовидную модель хранения метаданных документа в СЭД [3], что позволяет вносить дополнительные метаданные, не меняя структуры таблиц (рис. 6).

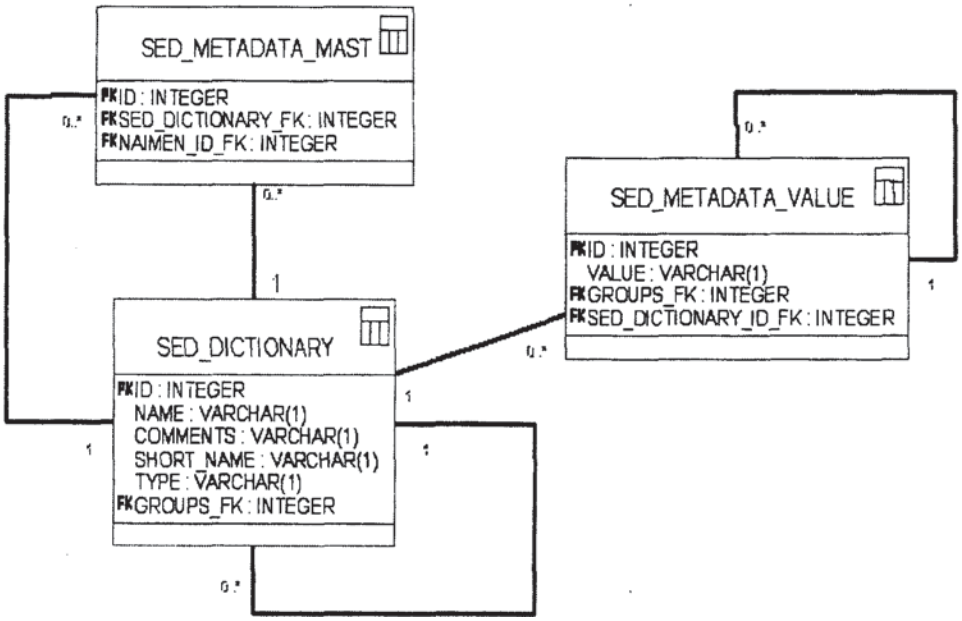

Рис. 6. ***ER-диаграмма метаданных документа в СЭД***

Помещение метаданных в реляционную базу данных связано с тем, что основная работа при поиске нужного документа, какой-либо информации, а также при взаимодействии систем друг с другом будет происходить с метаданными документа, и реляционная база данных как нельзя лучше подходит для этого, позволяя хранить и обрабатывать большое число записей, а также обеспечивать быстрый доступ к ним.

При реализации СЭД для создания функции интеграции с другими информационными системами, помимо использования стандарта для метаданных документа, необходим также стандарт представления данных. В качестве такового был выбран xml, на данный момент являющийся широко распространённым стандартом для представления данных при передаче их из одной системы в другую. Теги в xml создавались на основе описанных ранее метаданных.

Структура метаданных, построенная на основе общепризнанных стандартов, разработана таким образом, что позволяет как передавать описание документа по запросам другим информационным системам, так и стать основой для работы механизмов автоматизации документооборота вуза. Древовидная модель хранения позволяет вносить дополнительные метаданные, не меняя структуры таблиц, что изначально предусматривает возможность модернизации.

Компонент автоматического определения метаданных из электронного документа позволит упростить и ускорить процесс импорта документов в СЭД и станет важным компонентом системы интеграции подсистем в единую информационную систему.

Литература

1. Афанасьев, К. Е. Проблемы и типовые решения создания информационной инфраструктуры регионального образовательного комплекса / К. Е. Афанасьев, А. М. Гудов, Ю. А. Захаров, Б. П. Невзоров, И. В. Третьякова. – Кемерово: КемГУ, 2001.
2. Гудов, А. М. Об одной модели электронного документооборота вуза / А. М. Гудов, С. Ю. Завозкин, М. В. Семехина // Материалы VIII Международной конференции по электронным публикациям "El-Pub 2003".
3. Гудов, А. М. Система электронного документооборота / А. М. Гудов, С. Ю. Завозкин // XXV Всероссийская научно-методическая конференция КемГУ «Проблемы обеспечения качества образования». – Кемерово, 2004.
4. Метаданные. Описание. Обзор форматов метаданных // Информационно-интерактивный портал «Российские электронные библиотеки». http://www.rfbr.ru.
5. Dublin Core Metadata Element Set, Version 1.1: Reference Description. // http://dublincore.org.
6. ГОСТ Р 51141-98. Делопроизводство и архивное дело // http://mats.hotbox.ru/Year1/doc-gost-r5114198.doc.
7. ГОСТ Р 6.30-2003. Унифицированная система организационно-распорядительной документации // http://inform.alee.ru/docs/GOST_6_30-2003.pdf.
8. Суровцев, И. С. Нейронные сети / И. С. Суровцев, В. И. Клюкин, Р. П. Пивоварова. – Воронеж: ВГУ, 1994.
9. Уоссерман, А. Н. Нейрокомпьютерная техника: теория и практика / А. Н. Уоссерман. – М.: Мир, 1992.