От баз данных к базам знаний (объекты, формы, содержание)

В. Л. Арлазаров[1], Н. Е. Емельянов[2]

Рассматривается особая роль форм документов как основы взаимодействия с базами данных и знаний. Формы позволяют генерировать схемы баз данных, систематизировать расположение и обработку данных в СУБД, что существенно упрощает создание БД и аналитические исследования. Форма — структура из многих фреймов, элементарных порций знаний. Базы данных и знаний строятся на основе структуры объектов и понятий, представленных в схемах БД; шаблонов, макетов, форм отображения и процедур обработки знаний, представленных в формах, а также декларативных знаний, словарей и классификаторов, хранимых в СУБД.

Статья посвящена обсуждению роли форм (центрального понятия формулы: объекты — формы — содержание) как интерфейса с базами данных и знаний.

1. Немного истории

Коллектив программистов, сложившийся в конце 60-х гг. в Институте проблем управления РАН, затем в Институте системного анализа РАН, где активно работает до сих пор, уже более 30 лет постоянно занимается разработками систем управления базами данных (СУБД). Часть достижений стали классическими, такие как AVL деревья, базовый динамический метод доступа, индексация сложных структур данных, применение форм документов для генерации схем баз данных, организации ввода, распознавания, поиска, вывода, обработки данных, представление сложно структурированных данных в виде последовательностей (аналог XML, но широко внедренный раньше XML). Сегодня эти научные результаты используются многими другими разработчиками информационных технологий. Специфика созданных систем — сложные структуры объектов и подобъектов, возможность динамического изменения структур без перезагрузки БД. То

[1] 117312, Москва, просп. 60-летия Октября, д. 9, ИСА РАН, vla@cs.isa.ru.
[2] 117312, Москва, просп. 60-летия Октября, д. 9, ИСА РАН, nee@cs.isa.ru.

есть при необходимости пользователь может менять структуру и логику расположения данных в базе данных, что сложно или невозможно в СУБД других разработчиков. Это особенно важно при проведении исследовательской работы. Наши системы ориентированы на XML не только как на язык экспорта/импорта, но и как на аналог внутреннего представления данных. Созданы средства редактирования сколь угодно сложных и объемных документов, эффективные хранение, индексация и поиск XML документов, которые используют формы представления данных.

Этапы работ по созданию и развитию СУБД.

1978 г. Сдача 1-й версии СУБД ИНЭС (для ЕС ЭВМ) [1, 2]. Система была принята представительной Государственной комиссией: Академия наук, Министерства обороны, приборостроения, морского флота, культуры, Госплан, ЦСУ (на уровне зам. министров, начальников ГВЦ). Председателем комиссии был академик Г. С. Поспелов, возглавлявший тогда отечественные исследования по проблемам искусственного интеллекта.

1979 г. Система используется в 50 Министерствах и ведомствах, более чем в 100 ведущих организациях страны.

1982 г. Выход Постановления ГКНТ, рекомендующего всем Министерствам и ведомствам применять ИНЕС в АСУ и ИПС.

1982–1990 гг. Более 2000 ведущих организаций используют ИНЕС на более чем 50 000 рабочих местах.

1989–1995 гг. Разработка СУБД НИКА (аналог ИНЕС для IBM PC) [3] и около 500 поставок.

С **1995 г.** распространение систем Евфрат-Архив и Евфрат-Документооборот на СУБД НИКА — более 100 000 инсталляций.

Коллектив отмечен Премией Совмина СССР за разработку и массовое внедрение СУБД. На базе СУБД ИНЕС и НИКА были внедрены десятки тысяч крупных информационных систем.

2. Структурирование понятий и объектов предметной области

Построение базы данных или информационной системы об объекте управления или предметной области — это, по существу, описание знаний о них.

Как только начинаешь говорить об описании знаний, то невольно попадаешь в область гносеологии — главного раздела философии, которому больше 1 000 лет.

Рассмотрим высказывания философов о природе знаний:

«*Между чувственным образом и чистым познанием находятся схемы*».

Кант

«*Структура — это самое главное. Ведь без структуры нет раздельности. А если в предмете нет раздельности, то нет и свойств*».

А. Лосев

Разработчики СУБД быстро поняли важность схем, во всех СУБД появились языки описания схем БД. Второе высказывание тоже очень существенно, оно говорит о том, что объекты, понятия, если мы их хотим изучать, должны быть структурированы. Языки описания должны допускать введение сложных структур, так как иначе нет свойств, нет достаточно выразительных средств для описания содержания объектов — знаний о предметной области. Как человек может детализировать, уточнять любое понятие или вводить новые понятия обобщением введенных ранее, так и системы баз данных и знаний должны допускать это.

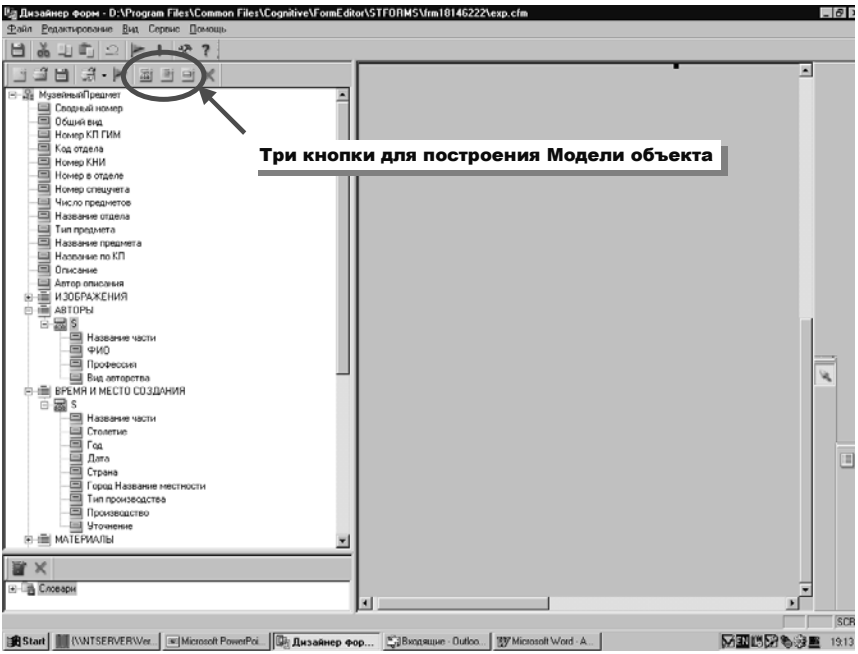

Рис. 1. Описание в Дизайнере форм модели объекта — содержания документа

Для описания структуры объекта оказывается достаточно трех кнопок (см. рис. 1): простое (неделимое) данное, структура (совокупность разнотипных данных), массив (совокупность однотипных данных). При помощи этих клавиш быстро строятся модели сложных объектов. В общем, сотни и тысячи реквизитов в описании объектов, которые определяют содержание документов, это нормально. Например, в паспорте музейного экспоната более 1 000 реквизитов.

3. Формы и фреймы для представления данных и знаний

Теперь обратимся к формам представления знаний. Начнем опять с философских высказываний.

> «*Содержание — определяющая сторона целого, совокупность его частей*». «*Форма — способ существования и выражения содержания*».
>
> БСЭ

> «*Содержание не бесформенно, а форма одновременно и содержится в самом содержании и представляет нечто внешнее ему*».
>
> Гегель

Другими словами: схема БД (модель содержания), это абстракция, отражающая взаимосвязи понятий. Со знаниями: декларативными и процедурными (факты и алгоритмы) можно работать только через формы. В случае БД это формы ввода/вывода, в случае программ — интерфейсы. В случае баз знаний — это и то, и другое.

Надо сказать, что формы с трудом входили в мир автоматизации программирования. Первые языки 1970-х гг. — COBOL (Common Business Oriented Language — ориентированный на задачи бизнеса) и RPG (Report Program Generator — генератор отчетов) не имели понятий «макет, бланк», «форма документа» — поразительно, но это факт [4]. Тогда было немало сильных программистов, которые доказывали, что формы не нужны, и без них все прекрасно[3]. Между тем уже в первой версии СУБД ИНЕС 1978 г. эти понятия были центральными.

[3] В «Англо-русском терминологическом словаре по банкам данных» (Минск, 1980) нет понятий документ и форма. Членам Рабочей группы по базам данных не удалось доказать в 1979 г., что это центральные понятия, которые чрезвычайно важны для организации интерфейса с БД. Словарь создавался вместе с американцами и довод против включения этих понятий в словарь был прост: американцы их не используют.

М. Минский предложил представлять знания как совокупность фреймов, его классическая работа опубликована в середине 70-х гг. [5]. Наши знания о мире, по мнению исследователей (вслед за Минским), представляют собой совокупность фреймов как структур данных для описания стереотипных ситуаций.

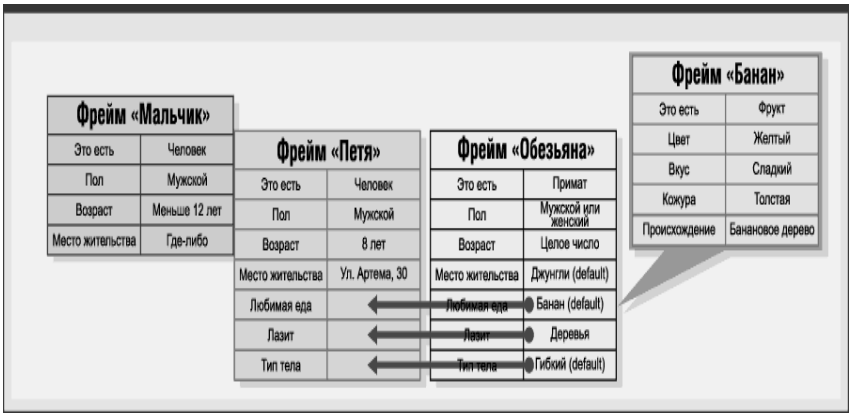

Рис. 2. Примеры фреймов

Информация, извлекаемая из опыта, хранится в памяти человека не в хаотическом беспорядке, а в виде разного рода связанных конструкций. В этом смысле фрейм используется для описания объектов, событий, ситуаций, прочих понятий и взаимосвязей между ними.

Фрейм — логическая запись, каждому полю (слоту) которой соответствуют основные элементы понятия. В формальных фреймовых моделях слотам ставятся в соответствие значения, присоединенные процедуры или другие фреймы.

С точки зрения специалиста по искусственному интеллекту (психолога, философа), фрейм — элемент знаний (порция знаний).

С программисткой точки зрения — это интерфейс работы со знаниями. Неважно, что там внутри (чтение из БД или выполнение какой-либо вычислительной процедуры). Главное, что фрейм имеет методы и интерфейсы:

- создание, корректировка, ввод/вывод порций разнородных знаний,
- выполнение запроса (например, типа QBE), который возвращает коллекцию объектов (порций знаний), удовлетворяющих условиям запроса,

- проверка непротиворечивости и целостности вводимых данных,
- запуск встроенных процедур (демонов) в случае того или иного события),
- запись в журналы операций работы с данными и т. п.

Можно сказать, что фреймы в исследованиях по искусственному интеллекту выполняют ту самую (философскую) роль способов существования и выражения содержания.

В контексте систем баз знаний полезно различать алгоритмические и неалгоритмические знания. Алгоритмические (или процедурные) знания — это алгоритмы (программы, процедуры), неалгоритмические — знания понятий и их взаимосвязей (понятийные знания) и фактов (декларативные знания). Мы будем следовать концепции, предложенной С. С. Лавровым (см. [6]. С. 278).

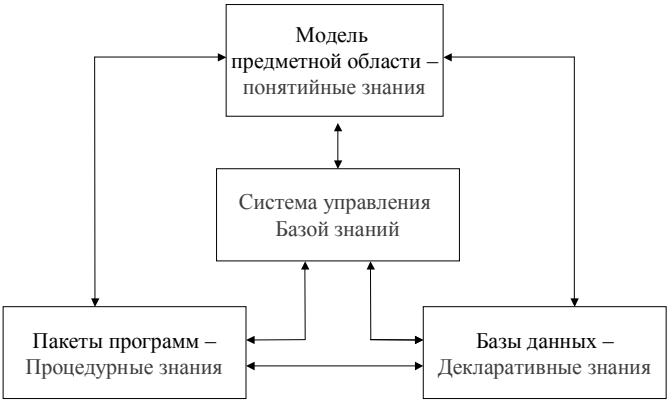

Рис. 3. Структура Системы управления базами знаний

На рис. 3 центральный блок «Система управления Базой знаний» представляет собой программу — «движок» (egine), который оперирует понятийными, процедурными и декларативными знаниями.

4. Формы документов

Если формализовать понятия *содержание* и *форма документа,* то можно доказать теорему двойственности [6]. Двойственность — математический термин, означающий, что по форме можно восстановить содержание и,

наоборот, по содержанию можно породить некоторую форму, позволяющую оперировать этим содержанием. Следовательно, если структура предметной области описана в виде схемы, то по схеме (или любой подсхеме) можно автоматически породить *форму* («*пустографку*») документа для ввода и отображения реквизитов и наоборот.

На рис. 4 показан пример формы (справа), автоматически построенной по структуре данных.

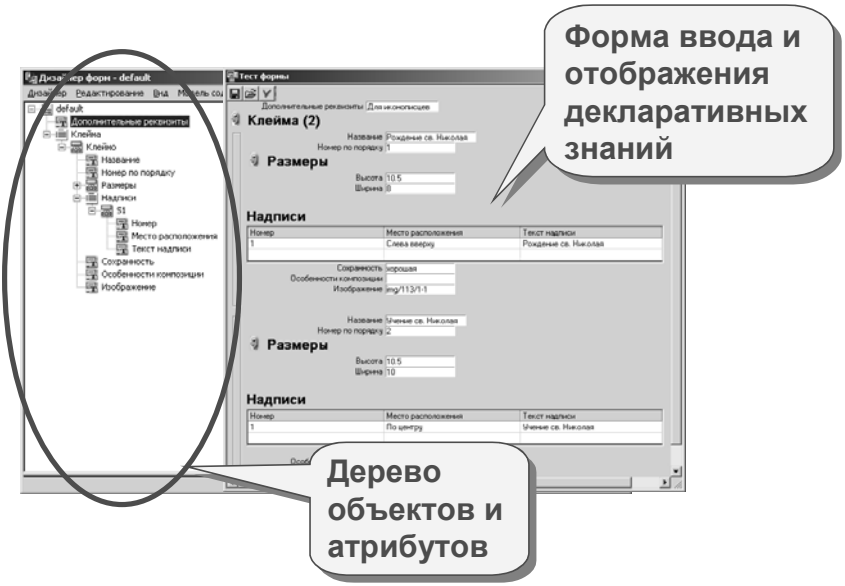

Рис. 4. Автоматическое создание формы по структуре данных

Форма очень емкое понятие: это не только окна для ввода информации, но и допустимые значения полей (словари), они на рис. 5 представлены в левом нижнем углу. Их обычно много — сколько полей ввода. Каждое поле (слот фрейма) имеет много спецификаций — они представлены справа вверху, каждое данное тоже имеет много спецификаций — справа внизу.

Большие сложно структурированные формы образуют иерархии фреймов. Так, на рис. 6 представлена форма описания экспоната музея с более чем 1 000 реквизитами, разбитая примерно на сто фреймов, выбор представленного на рисунке фрейма получается выбором двух закладок «Основные данные» и «Описание».

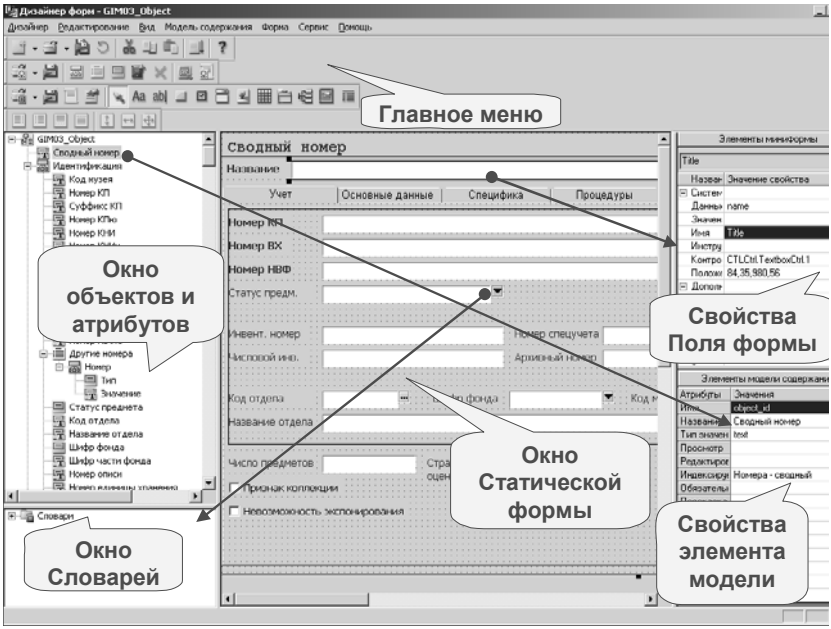

Рис. 5. Компоненты формы. Окна дизайнера форм

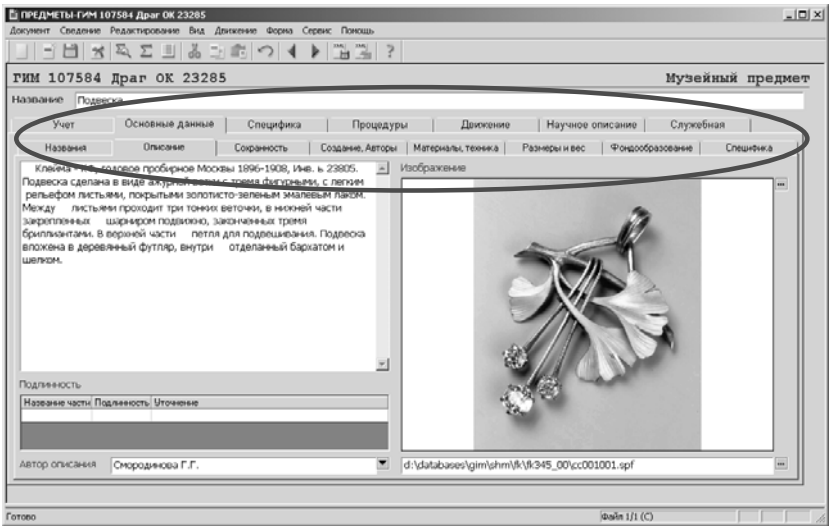

Рис. 6. Форма описания объекта — иерархия фреймов

К формам относятся также и процедурные знания — алгоритмы вычисления значений полей, проверки правильности заполнения полей и др. На рис. 7 представлено описание поля «Сводный номер», которое вычисляется по определенному алгоритму (процедурному знанию).

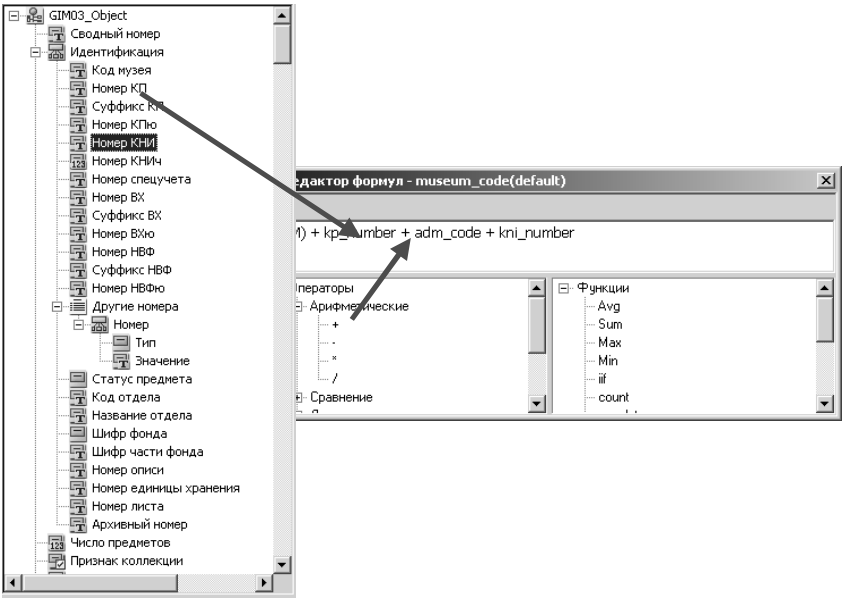

Рис. 7. Описание процедур, связанных с формой

Изложенные выше принципы представления знаний в виде форм реализованы в программном комплексе x-Ника технология [8]. Технология названа x-Ника потому, что в ее основе лежат XML-документы и объектно-ориентированная СУБД НИКА — наследница ИНЕС. Некоторые ее характерные элементы кратко рассмотрим в следующем пункте.

5. x-Ника технология построения информационных систем по формам ввода/вывода

Спецификой x-Ника технологии является возможность построения информационных систем по формам входных и выходных документов. По

комплекту входных форм генерируется схема БД и структура объектов создаваемой информационной системы с их взаимосвязями. Основой х-Ника-технологии является широкое использование формата XML [9]. Это и многоуровневые коллекции объектов, и хранилище (типа XML DB [10]) сложно структурированных информационных объектов (описаний экспонатов), построенное на СУБД НИКА, которая обеспечивает хранение объектов любой структурной сложности. Кроме того, СУБД НИКА дает возможность рассматривать объекты как многоуровневые коллекции подобъектов.

Используется также основанный на XML стандарт представления форм входных и выходных документов и собранных по этим формам данных, включая изображения и т. п. Так как сейчас практически все промышленные СУБД содержат средства вывода данных в формате XML, это свойство позволяет довольно просто решать задачи смены платформы, например, переход с Access на х-Ника технологию.

При этом понятийные знания хранятся в схеме БД, процедурные знания в формах, декларативные знания в СУБД. Создание информационной системы осуществляется средствами двух дизайнеров: дизайнера форм (экранных форм для ввода и отображения данных), средствами которого описываются как струтуры данных, так и процедуры обработки, и дизайнера отчетов (форм вывода данных), который также содержит средства описания процедур обработки, необходимые для формирования отчетов.

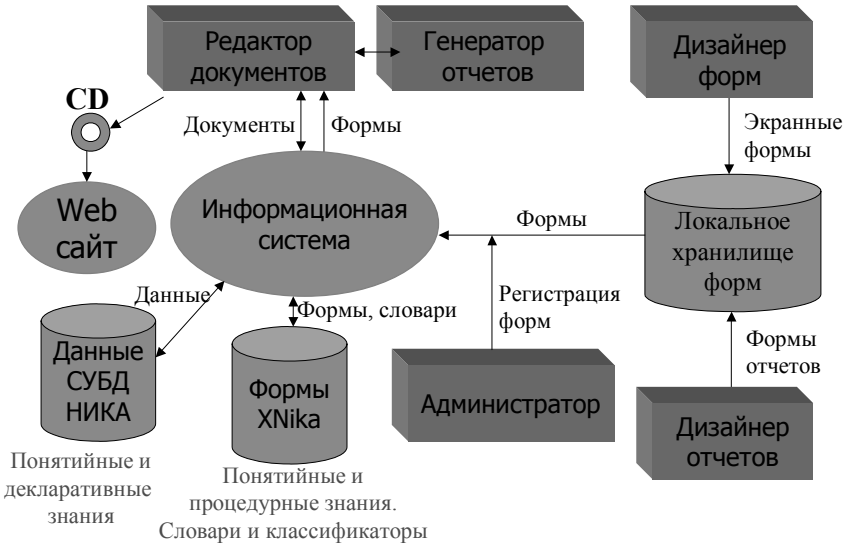

Рис. 8. Архитектура х-Ника технологии

На рис. 8 представлена архитектура х-Ника технологии. Дизайнеры форм входных и выходных документов помещают их в хранилище форм. Система Администратор позволяет создавать ИС по указанным входным и выходным формам, определяет пользователей и их права. Редактор документов служит для ввода, поиска и просмотра документов. Данные сохраняются в БД. Система CD-maker (изготовитель CD) — позволяет создать компакт диск или программу для сенсорного киоска. Работа с CD и киосками осуществляется средствами стандартных браузеров. Этот же пакет можно открыть в Интернете.

Любая информационная система создается автоматически по формам входных и выходных документов. Порождается минимальная схема БД, содержащая все объекты и реквизиты входных форм. Создается индекс по всем (по умолчанию) реквизитам всех объектов. Система готова для ввода и отображения данных по любому нерегламентированному запросу конечного пользователя, а также для печати выходных документов. Технология позволяет создавать по любому фрагменту БД, получаемому в результате выполнения запроса, отторгаемые информационные ресурсы на CD, DVD, информационных киосках. Эти ресурсы, которые включают СУБД и Web сервер, можно располагать на интернет-порталах.

5. Заключение

Рассмотренные выше технологии успешно применялись для построения широкого круга задач. Примеры применений: системы документооборота, учет экспонатов музея, библиотечные системы, системы отчетности для учебных заведений, описание уникальных научных аппаратов, истории болезней, ведение архивов (фотографии, документы), описание видов новейшего вооружения и т. п.

Редактор сложно структурированных документов и хранилище документов на основе СУБД НИКА являются, по существу, средствами подготовки, хранения, визуализации и корректировки широкого класса XML документов. Методы и средства работы с XML документами сейчас активно разрабатываются научными и коммерческими организациями во всем мире. Самый популярный сейчас редактор XML документов — XML Spy фирмы Altova [11]. Он имеет следующие недостатки: не работает с формами, редактируемый документ должен быть в оперативной памяти, нет коллекций документов, каждый документ — отдельный файл. Необходим более широкий подход к XML как к языку представления знаний. х-Ника технология — шаг в этом направлении.

Важной представляется разработка единой формы для всех видов когнитивной трансформации данных: распознавание и ввод, отображение на

экране и редактирование, печать документов, запросы и др. Первая реализация такой формы, включающей все спецификации разных видов использования форм, представлена в [11].

Литература

1. *Арлазаров В. Л., Емельянов Н. Е. и др.* Информационная система ИНЭС // Автоматика и телемеханика. М., 1979. № 6.
2. *Емельянов Н. Е.* Введение в СУБД ИНЕС. М.: Наука, 1988. С. 256.
3. *Годунов А. Н., Емельянов Н. Е. и др.* Система НИКА // Системы управления базами данных и знаний. М.: Финансы и статистика, 1991. С. 209–248.
4. *Арлазаров В. Л., Емельянов Н. Е., Жаринов А. Н.* Сравнительное описание программных средств вывода // Алгоритмы и организация решения экономических задач. Сб. статей. Вып. 10. М.: Статистика, 1977. С. 59–69.
5. *Минский М.* Фреймы для представления знаний. Киев, 1979 (Первое издание на русском языке).
6. Системы управления базами данных и знаний. Справ. изд. / М.: Финансы и статистика, 1991. С. 352.
7. *Bogacheva A. N., Emelianov N. E.* Duality between Document Structure and Data Base Structure // Proceedings of the Workshop on Advances in Database and Information Systems. ADBIS' 94. M. May 1994. P. 83–89.
8. *Богданов А. С., Емельянов Н. Е., Ерохин В. И. и др.* НИКА — технология построения информационных систем // Организационное управление и искусственный интеллект / Сб. трудов ИСА РАН. Под ред. члена- корр. РАН Арлазарова В. Л. и д-ра. тех. наук Емельянова Н. Е. М.: УРСС, 2003. С. 52–67.
9. XML (The Extensible Markup Language) (см.: http://www.w3.org/TR/REC-xml).
10. *Грейвс М.* Проектирование баз данных на основе XML. Изд. дом «Вильямс», 2002. 640 с.
11. XML Spy (см. http://www.altova.com/products_ide.html).
12. *Арлазаров В. В.* Структурирование визуальных представлений информационной среды и методы определения надежности распознавания. Автореф. дис. ... канд. тех. наук. М., 2004. С. 25.