

**КОМПОНАТ ИССЛЕДОВАТЕЛЬНОЙ И
ЭКОНОМИЧЕСКАЯ КОМИССИЯ ДЛЯ ЕВРОПЫ**

**КОМИССИЯ
ЕВРОПЕЙСКИХ СООБЩЕСТВ**

КОНФЕРЕНЦИЯ СТАТИСТИКОВ ЕВРОПЫ

ЕВРОСТАТ

**Совместная рабочая сессия ЭКЕ и ЕВРОСТАТА
по электронному представлению данных
(Женева, Швейцария, 13 - 15 Февраля 2002 г.)**

Рабочий доклад №16

Тема 3: стандарты и концептуальные модели метаданных

МОДЕЛЬ МЕТАДАНЫХ "IQML" ДЛЯ СБОРА ИНФОРМАЦИИ

Заказной доклад

Представлен Джоанной Лэм, Эдинбургский университет, Великобритания

Резюме: В этом докладе описывается проект "IQML" (комплект программных средств и стандарт языка XML для самонастраивающегося опросного листа), который осуществляется под руководством Евростата в рамках Пятой Рамочной программы Европейского Союза по научным исследованиям и технологическим разработкам. "IQML" состоит из пяти связанных программных модулей, имеющих общую модель данных. При этом, определенные модули используют свою собственную модель, которая не только соответствует общей модели, но и обладает дополнительными возможностями, необходимыми для выполнения определенных функций. В докладе представлены общая модель, связанные модели и установление соответствия между ними. Кроме того, описывается деятельность комитета проекта "IQML" по распространению данных моделей среди тех, кто пользуется соответствующими стандартами для разработки программного обеспечения и в сфере официальной статистики.

I. ВВЕДЕНИЕ

1. "IQML" (IST-1999-29093) - это проект с распределенными затратами, который является частью Пятой Рамочной программы Европейского Союза. Среди участников проекта три коммерческие компании ("Dimension EDI", "Comfact AB" и "DESAN Marktonderzoek"), два государственных статистических ведомства (CSO, Ирландия и Статистика Норвегии) и два университета: Афинский государственный технический университет и Эдинбургский университет, который координирует ход проекта. Проект начался в феврале 2000 г. и закончится в январе 2003 г.

2. В "IQML" входят следующие программные модули:

- Инструментарий представления опросного листа, разработанный "Comfact AB"
- Инструментарий запроса базы данных, разработанный "Comfact AB";
- Инструментарий дизайна опросного листа (ИДОЛ), разработанный Эдинбургским университетом,
- Инструментарий управления опросом, разработанный "DESAN",
- Репозиторий метаданных, разработанный "Dimension EDI".

3. Репозиторий играет две роли: в нем хранится центральная модель метаданных "IQML", а кроме этого он достаточно обобщенный, чтобы дать возможность другим модулям хранить их объекты, хотя в этом случае репозиторий более ограниченно поддерживает хранимый объект. Инструментарий представления опросного листа и инструментарий запроса базы данных используют общую модель, ориентированную на сбор данных. Инструментарий дизайна опросного листа имеет модель, ориентированную на создание анкет и схожие исследовательские

инструментальные средства. Инструментарий организационного управления опросом имеет модель, направленную на координирование работы администрации и мониторинг опросных листов различного типа (бумажных, электронных, сетевых и т.д.). Модели разработаны на UML [1] (едином языке моделирования) и выполнены на языке XML. Роль репозитория состоит в том, чтобы получать метаданные, используемые всеми модулями, а также чтобы хранить метаданные, необходимые для других задач процесса управления данными, таких как редактирование, анализ и распространение.

4 Модули не только обеспечивают полный комплект программных средств для сбора статистических данных посредством опросных листов и анкет, они также функционируют как независимые пакеты, а репозиторий обеспечивает открытое согласование, которое даст возможность другим программам взаимодействовать с модулями "IQML".

II МОДЕЛЬ "IQML"

5 Модель "IQML" предназначена для фиксирования основных аспектов процесса разработки и управления опросом. В рамках жизненного цикла статистического процесса, она охватывает разработку опросного листа, управление опросом, сбор данных и хранение связанных с ними метаданных в доступном формате. Модель не охватывает процесс индексации, анализ и распространение результатов статистических исследований.

6 Модель представлена графическим языком под названием UML (единый язык моделирования), который может быть реализован посредством использования одной из многих систем программного обеспечения, предназначенных для практического применения этого языка и предоставляющих его описание. Для этой задачи в комплекте "IQML" используется программный пакет "Rational Rose" [2].

7 В данном докладе для простоты изложения мы опишем модель текстуально. Поскольку репозиторий метаданных является основным средством передачи метаданных от отдельных моделей всей системе в целом, с него мы и начнем. При первом появлении в тексте терминов, используемых в модели "IQML", они будут даны курсивом.

8 Репозиторий метаданных (PM) состоит из четырех основных пакетов: "*Базиса*" (*Foundation*), "*Классификации*" (*Classification*), "*Банка вопросов*" (*Question Bank*), и "*Административных метаданных*" (*Administration Metadata*). Пакет "Базис" имеет две основные ориентации: содержание и формулировка. Под "*Содержанием*" (*Content*), мы имеем в виду приглашения, которые содержатся в опросном листе. Обычно это текст, однако графика, а в web-анкетах анимация и звук тоже могут быть квалифицированы как "содержание". "Текстовая" часть опросного листа может представлять из себя смесь слов и графики, таким образом упорядоченное множество "Содержания" может храниться как "*Группа содержания*" (*Content Group*).

9 В соответствие со своей структурой опросной лист запрашивает "Содержание" в нескольких различных позициях, обобщенное название всех этих позиций - "*Узел*" (*Node*). Различные типы "Узлов" включают вопросы, секции, переменные величины, документы и т.д. "Группа содержания" контактирует с узлом через "*Контекст*" (*Context*). "Контекст" определяет обстоятельства, в которых узел может быть размещен. К "Контексту" относятся, например, язык, определение популяции, средство доставки.

10 Объекты в пакете "Базиса" обязательно являются чем-то абстрактным. На Рисунке 1 дана схема, иллюстрирующая все эти взаимоотношения.

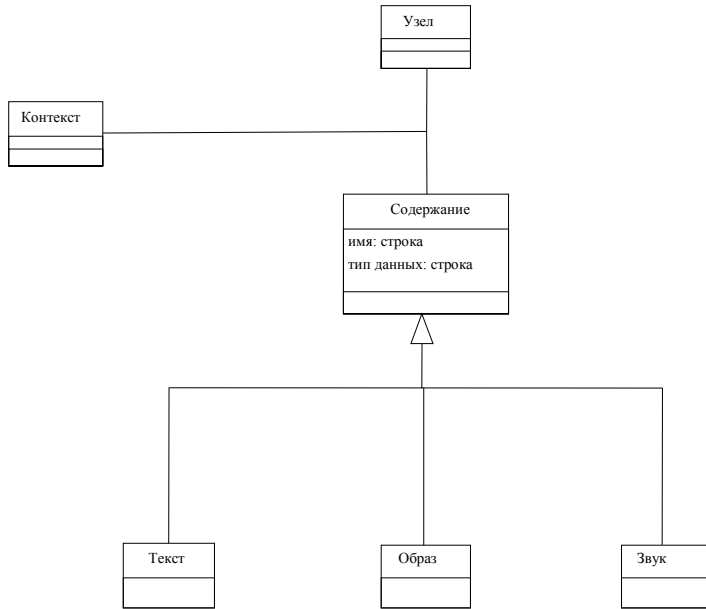


Рисунок 1

11 Пакет "Классификации" также является общим пакетом, поддерживающим все иерархические структуры. По этой причине, модель является рекурсивным представлением "Узлов" и листьев иерархического дерева. Более детально "Классификации" обсуждаются в разделе VII.

12 Пакет "Банк вопросов" содержит "Вопросы" (*Questions*) и "Секции" (*Sections*), которые могут быть использованы в одном или нескольких опросных листах. "Секция" представляет из себя упорядоченное множество "Вопросов". Таким образом "Вопрос" может быть частью "Секции" или отдельным. "Вопросы" и "Секции" являются разными типами "Узла". Опросной лист устроен таким образом, что часто "Концепт" (*Concept*) определяется раньше, чем формулируется подробный вопрос. "Банк вопросов" может хранить иерархические наборы "Концептов". "Концепты" связаны с "Вопросами", а листья дерева "Концептов" могут относиться к одной "Переменной" (*Variable*). "Переменная" также является подклассом "Узла". Это основная концепция "Банка вопросов", проиллюстрирована на Рисунке 2.

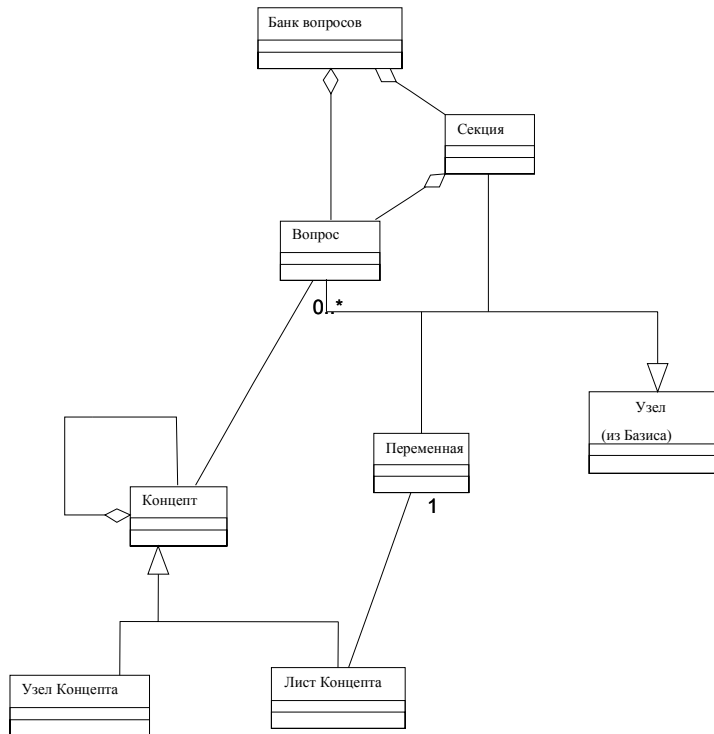


Рисунок 2

13 Опросной лист создается посредством обращения по ссылке, к "Вопросам" и "Секциям" из "Банка вопросов". Опросной лист может иметь "Контекст", определяющий, какой вариант "Содержания" должен быть связан с определенными "Вопросами" и "Секциями".

14 Модель также дает элементы управления навигацией, состоятельностью данных и вычислениями. "Правило процесса" (Process Rule) ассоциировано с "Узлом". "Правило процесса" - это алгоритм, выдающий оценку "ПРАВИЛЬНО" или "НЕПРАВИЛЬНО". С "Правилом процесса" ассоциировано одно или несколько "Действий" (Actions). Есть два типа "Действий": "Правильное действие" (IfTrueAction) и "Неправильное действие" (IfFalseAction). Соответствующее "Содержание" связано с "Узлом", согласно спецификации "Действия".

15 Инструментарий организационного управления опросом касается деталей индивидуализированных объектов, и поэтому основная обработка происходит на уровне данных, а не на уровне метаданных. Тем не менее определенные виды ключевой информации необходимо зарегистрировать для последующей обработки статистических данных. Прежде всего, следует упомянуть подмножество "Переменных", именуемое "Административные переменные" (AdminVariables). Они используются для принятия решений, связанных с обработкой данных опроса. Затем мы обнаружим три подкласса этой переменной: "Административный номер" (AdminNumber) однозначно определяет единицу выборки. "Административный компонент" (AdminComponent) - это переменная, посредством которой создается "Административный номер". Следует подчеркнуть, что метаданные не дают никакого уникального идентификатора; они просто определяют переменную, в которой это хранится и устанавливает алгоритм, необходимый для получения номера. Другие переменные, используемые в процессе управления, хранятся в остаточном классе. Обычно это элементы подвыборки, средства, используемые для сбора данных и другие переменные, ориентированные на контекст. Регулируемые переменные, как, например, дата отправки и алгоритм напоминания, тоже относятся к категории, известной как "Компонент".

выводимой записи" (*OutputRecordComponent*). "Выводимая запись" (*OutputRecord*) - это коллекция всех "Компонентов выводимой записи". "Спецификация выводимой записи" (*OutputRecordSpecification*) определяет "Компонент выводимой записи". "Запускающие события" (*EventTriggers*) определяют некоторые из "Спецификаций выводимой записи", эти события подразделяются на "События ввода" (*InputEvents*) и "События вывода" (*OutputEvents*). Описанные концепции иллюстрируются на Рисунке 3.

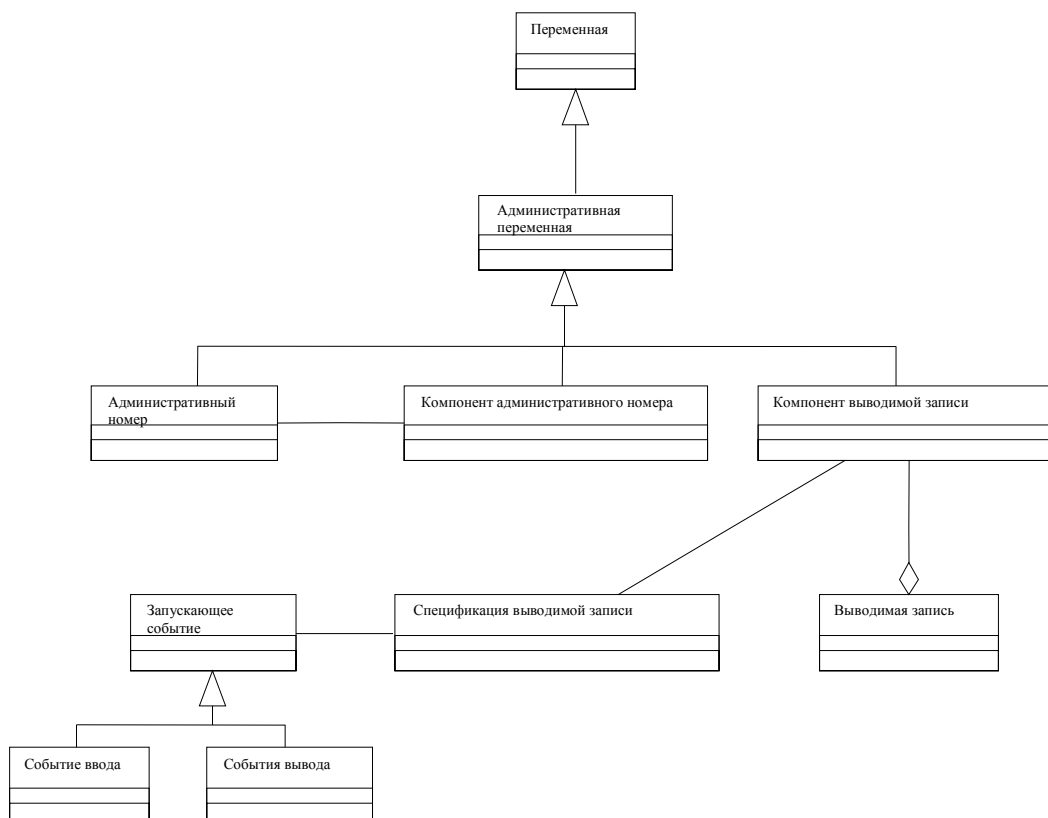


Рисунок 3

III МОДЕЛЬ ПРЕДСТАВЛЕНИЯ

16 Инструментарий представления опросного листа (ИПОЛ) используется для представления опросных листов на языке XML в Web-браузерах на ПК для ручного ввода данных. Он обеспечивает возможность проверки беспристрастности выборки, навигации и вычислений. Инструментарий использует такие доступные Web-браузерные технологии, как DOM, DHTML, CSS и Java Scripts. Механизмы защиты и конфиденциальности информации интегрированы и используют стандартные продукты. На Рисунке 4 показаны компоненты Инструментария представления.

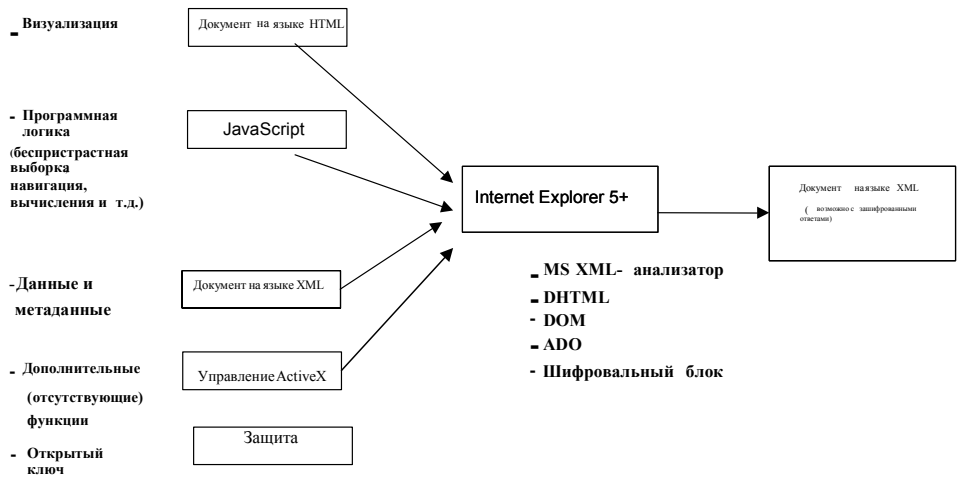


Рисунок 4

1.

17 Визуализация достигается посредством использования языка HTML. Для выполнения операций программной логики, то есть для управления беспристрастной выборкой, навигацией и вычислениями, используется JavaScript. Данные и метаданные содержатся в документе на языке XML. Технология "ActiveX" используется для придания системе дополнительных функциональных возможностей. Для обеспечения системы защитными функциями в нее интегрируется имеющееся программное обеспечение.

18 Модель данных у инструментария запроса базы данных (ИЗБД) такая же, как и у ИПОЛ. Запрос базы данных обеспечивает полуавтоматическое извлечение информации из одной или нескольких ODBC-совместимых баз данных. Отдельный XML-файл содержит отображение элементов данных опросного листа к одной или нескольким базам данных. Восстановленные данные сохраняются в XML-опроснике того же формата, что и ИПОЛ. ИЗБД почти не нуждается в дополнительных программах, за исключением Web-браузера (IE5+) и использует такие технологии, как Java Scripts, ODBC, DOM, и т.д. Для работы с ним требуется знание структуры местной базы данных и SQL (языка структурированных запросов).

может включать в себе заранее заполненную информацию.

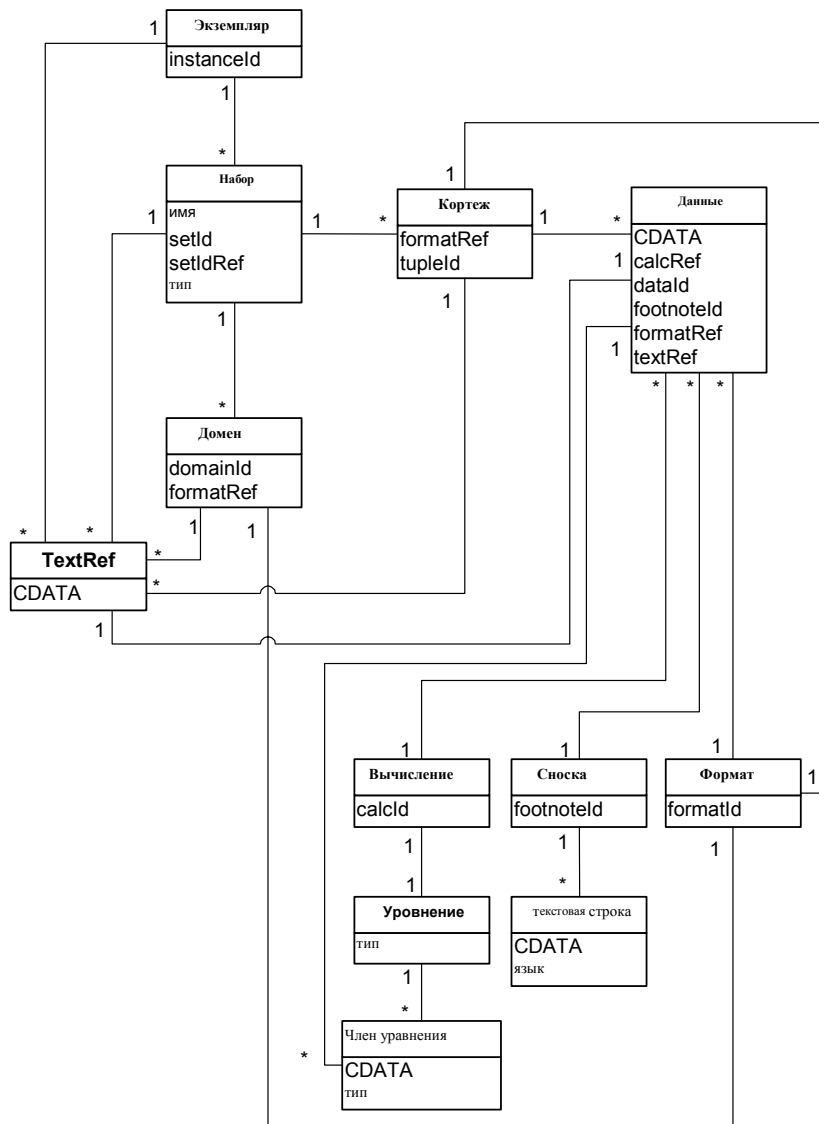


Рисунок 6

IV ДИЗАЙНЕРСКАЯ МОДЕЛЬ ОПРОСНОГО ЛИСТА

21 Дизайнерская модель опросного листа дает возможность разработчику создать опросный лист, который соберет необходимые данные. В зависимости от типа респондента, эта задача может быть или очень простой, или требующей весьма высокой квалификации. Можно выделить три стадии в процессе разработки опросного листа. Прежде всего, необходимо определить характер опроса и выяснить, каковы нормативы показателей. Во-вторых требуется отобрать подходящие типы вопросов и уметь четко и ясно их сформулировать. В-третьих, нужно выстроить логичную последовательность сформулированных вопросов. Поскольку эти стадии повторяются при создании каждого вопросника, навыки разработчиков и качество анкет постоянно совершенствуются. Разные разработчики по-разному подходят к решению этой задачи, поэтому каждая из вышеназванных стадий может быть детально продумана по ходу работы над вопросником. В дизайнерской модели опросного листа есть "Банк вопросов" (*Question Bank*), в котором хранятся "Вопросы" (*Questions*) и группы "Вопросов" ("*Секции*" [*Sections*]),

представляющие собой неупорядоченный набор. "Вопросник" (*Questionnaire*) - это упорядоченный список ссылок на "Вопросы" и "Секции" в "Банке".

22 С "Банком" связан метафайл ИДОЛ, определяющий метаданные для "Банка вопросов". Два типа метаданных используются для определения того, как должны выглядеть вопросы, и какого рода данные они должны собрать. "*Tun voprosa*" (*Question Type*) классифицирует вопрос в соответствии с тремя критериями: тип данных, тип ответа и использование подвопросов. Эта концепция проиллюстрирована на Рисунке 7.



Рисунок 7

23 "Вопросы" квалифицируются, в зависимости от типа данных (целое число, текст, булево выражение (да/нет) и т.д.), типа ответа (простой, одновариантный, многовариантный), а также в зависимости от того, есть ли у "Вопроса" "подвопросы". Если есть, вопрос выглядит в виде таблицы или сетки. Типы вопросов определяются пользователем, поэтому эта терминология вполне пригодна для внутреннего пользования. Легко заметить, что эта структура очень близко воспроизводит схему ИПОЛ.

24 Шесть типов "*Текстовых элементов*" (*Text Element*) связаны с опросным листом: "Вопросник", "Секция", "Вопрос", "Ответ", "Подвопрос" и "Дополнение". Эти элементы могут быть разбиты пользователем на подклассы таким образом, чтобы по всему опросному листу сохранялось стилевое единство. Например, нижеприведенный вопрос имеет два подкласса: верхнюю строку, выделенную жирным шрифтом, и дополнительный текст в скобках.

Общее количество занятых на предприятии в дни недели, заканчивающейся 17 июня 2001 г.
(включая работающих по найму, директоров, контрагентов и членов семьи)

V МОДЕЛЬ ОРГАНИЗАЦИОННОГО УПРАВЛЕНИЯ

25 Модуль организационного управления опросом - это инструментарий проведения обследования. Он связывает вопросник из Инструментария дизайна опросного листа с выборкой и контролирует процесс отсылки и возвращения вопросника. На Рисунке 8 показаны основные объекты модели. Объекты, выделенные жирным шрифтом, относятся к деятельности, связанной с проведением обследования, а остальные хранятся на уровне индивидуализированного объекта. Инструментарий организационного управления опросом дает возможность руководителю производить выборку и связывать ее единицы с опросными листами. Объект "*Управление*

откликом" (*SampleUnitResponseAdmin*) содержит информацию о реакции респондента (вернул опросные листы, нет связи, отказался и т.п.). После завершения опроса эта информация может быть привязана к определенному опросному листу.

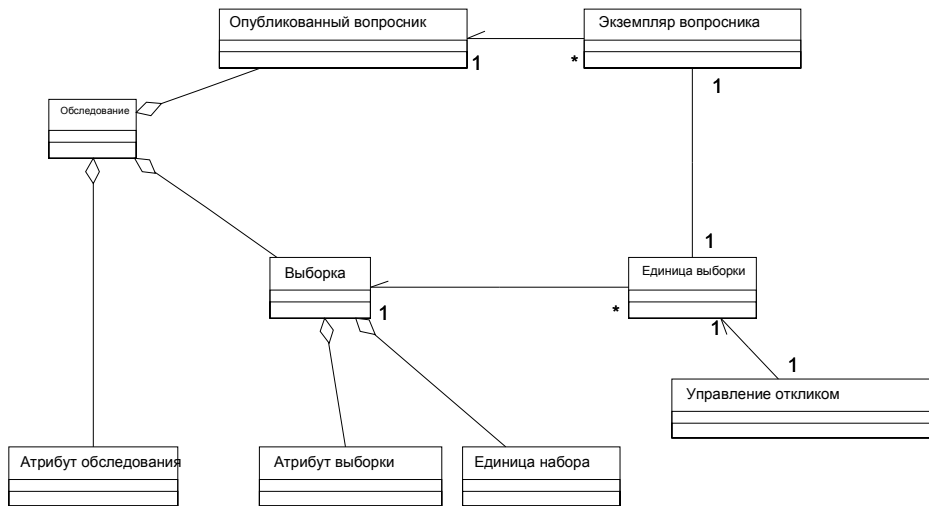


Рисунок 8

VI РЕПОЗИТОРИИ МЕТАДААННЫХ

26 Репозиторий метаданных используется для хранения метаданных модели "IQML", что необходимо для того, чтобы дать возможность обмена метаданными между инструментариями, а также для того, чтобы метаданные были доступны для инструментов третьей стороны. Репозиторий независим от модели, которая определяется для репозитория при запуске. Соответственно, репозиторий будет отклонять объекты, которые не соответствуют модели. На Рисунке 9 показаны объекты в репозитории. Модель представляет собой упрощенную версию модели ebXML [3].

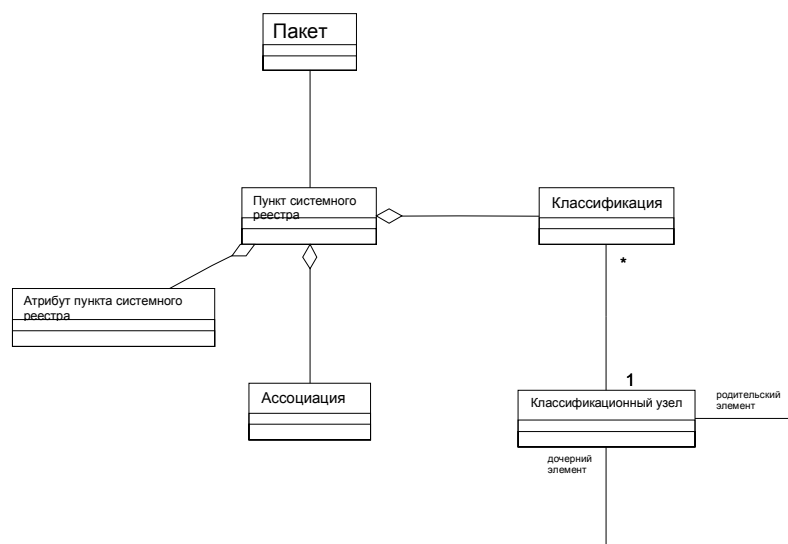


Рисунок 9

VII СВЯЗЬ И СООТВЕТСТВИЯ

27 Рисунок 10 иллюстрирует функциональную связь, существующую между модулями проекта. Репозиторий сконфигурирован таким образом, что информация может использоваться всеми приложениями. Возможно также выделить из репозитория язык XML "IQML" для использования ИПОЛ и ИЗБД. Мы отделяем специальные метаданные приложения, которые хранятся в локальном запоминающем устройстве, от модели "IQML", которая хранится в репозитории. В настоящее время определенные части модели расширяются, чтобы придать им больше функциональных возможностей.

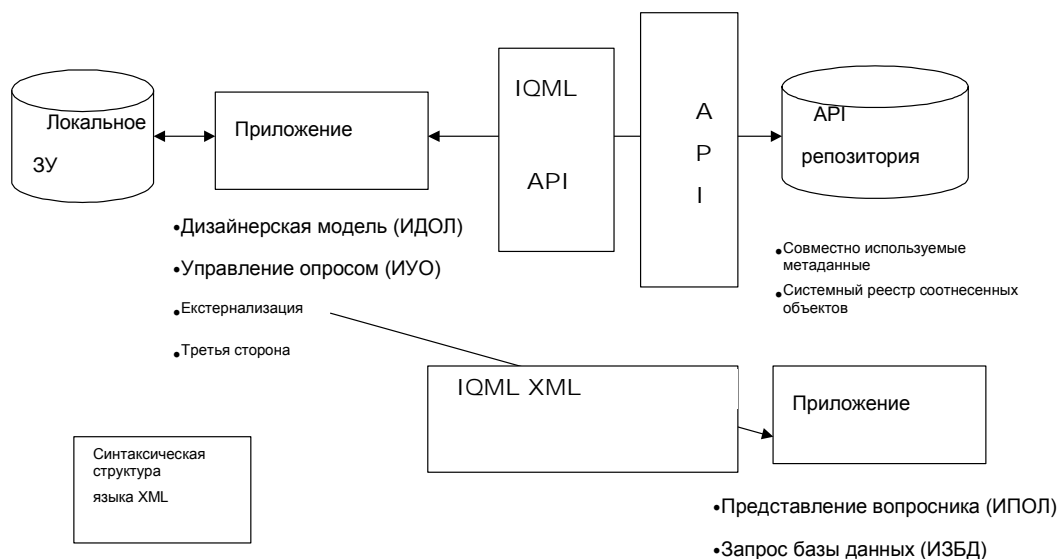


Рисунок 10

VIII СТАНДАРТЫ И ОБМЕН

28 "IQML" XML был принят EEG6/WG4 в качестве основы для RDRMES/XML. XML-документы, соответствующие этому стандарту, могут быть извлечены непосредственно из репозитория метаданных. Кроме того, программное обеспечение третьей стороны может извлечь метаданные из репозитория, чтобы адаптировать их к своим стандартам. Мы работаем над заявкой на предложения (RFP) в области опросных листов и сбора данных в составе специальной подгруппы, занимающейся проблемами аналитического управления данными и входящей в состав рабочей группы по развитию стандартов объектного программирования (OMG¹). С подачей RFP-заявки официально начинается процесс утверждения спецификации в качестве стандарта.

29 Работа над проектом "IQML" проходит не в изоляции. Помимо деятельности, описанной выше, партнеры вносят свой вклад в развитие сети "MetaNet" [5]. Эта сеть относится к группе "COSMOS" проекта "IST", который осуществляется в рамках Пятой Рамочной программы ЕС [6]. В результате модели, которые мы разрабатываем, будут использоваться и всеми другими заинтересованными в моделировании метаданных участниками упомянутых проектов.

¹ <http://www.omg.org/>

IX ЗАКЛЮЧЕНИЕ

30 "IQML" - это трехлетний научно-исследовательский проект, который завершится в январе 2003 года. К настоящему времени мы разработали базовую модель и внедрили первые опытные образцы всех модулей (и второй опытный образец ИПОЛ). Оценочные испытания первого опытного образца назначены на декабрь 2001 г. В течение заключительного года проекта будет продолжено совершенствование и окончательно завершена разработка общей модели, описанной в разделе II. "IQML" вносит весомый вклад в развитие моделей и программного обеспечения для сбора первичных данных тремя путями. Во-первых, мы разработали XML-схему, которая будет использована в качестве основы для логического блока передачи первичной информации по стандарту EEG6/WG4. Во-вторых, у нас будет полный набор программного обеспечения для сбора первичных данных и хранения связанных с этой задачей метаданных. В-третьих, мы вносим свой вклад в создание общей эталонной модели статистических метаданных, сотрудничая с другими участниками проекта в области сбора данных.

Ссылки

- [1] Booch, G и др., "Руководство пользователя универсального языка моделирования", 1999 г., The Addison-Wesley Object Technology Series, ISBN 0-201-57168-4
- [2] Программный пакет "Rational Rose 2000", корпорация "Rational Software", <http://www.rational.com/index.jsp>.
- [3] Webber, D. и Kotok, A. "ebXML: Новый всемирный стандарт для Интернет-коммерции", 2001г., издательство "New Riders Publishing", ISBN: 0735711178, <http://www.newriders.com>
- [4] OMG, "Спецификация CWM (общей метамодели хранилища данных)", 2000 г., Документ OMG от 01.01.2000 г., 11 февраля 2000 г.
- [5] Lamb, J.M. "MetaNet, непревзойденная сеть для гармонизации и синтезирования развития статистических метаданных", 2000 г., CES, Эдинбургский университет, <http://www.epros.ed.ac.uk/metanet/>
- [6] Lamb, J.M. "COSMOS: пакет систем метаданных для официальной статистики", 2001 г., CES, Эдинбургский университет, <http://www.epros.ed.ac.uk/cosmos/>