УДК 681.324

Шахгельдян К.И.

Проблемы качества данных и информации в корпоративной информационной среде вуза

Аннотация

В работе рассматриваются проблемы, связанные с обеспечением качества данных и информации в сложной корпоративной информационной среде вуза. Для автоматизации процедур поддержки качества предлагается использовать обобщенный репозиторий метаданных информационной среды.

Введение

Стратегические цели информатизации вуза связаны с повышением эффективности управления и качества образования одновременно со снижением риска принятия необоснованных решений [1]. Корпоративная информационная среда (КИС) является основным инструментом процесса информатизации вуза. Цели информатизации определяют повышенное внимание к вопросам качества данных и информации в КИС.

В работе [2] информационная система рассматривается как проекция систем реального мира на область информационных технологий (ИТ). Проекция должна отображать любое состояние системы реального мира в тот же или иной момент времени. КИС можно рассматривать как объединение информационных систем, с дополнительными требованиями, обусловленными, во-первых, сложностью КИС, во-вторых, необходимостью управлять процессами организации, используя КИС.

Понятие сложных систем можно найти в общей теории сложности, например, в работах [3]. В сложной КИС поведение определяется не только поведением отдельных систем, но и их взаимодействием. В этой связи интересен вопрос получения качественной информации в КИС, так как информация, полученная в отдельной системе КИС, может противоречить информации полученной в другой системе КИС. Сложная КИС может быть

неустойчивой, т.е. небольшие нарушения в качестве данных в одной системе КИС могут привести к катастрофическим нарушениям качества в другой. Важным элементом сложной КИС с точки зрения качества данных и информации является так же и постоянные изменения, происходящие в сложных системах, поэтому обеспечение качественной информации может быть корректным только на текущий момент времени, и нарушаться с любым изменением в КИС.

Определим понятие информационной среды как сложной информационной системы (или совокупности информационных систем), которая является активной проекцией реального мира на область ИТ, т.е. информационная среда влияет на системы реального мира. Иначе говоря, информационная среда – это сложная информационная система (или совокупность информационных систем) с обратной связью.

В классических работах по качеству данных и информации [4-8] обсуждаются проблемы качества в информационных системах. В этой работе мы рассмотрим проблемы качества данных и информации в КИС.

1. Качество информации и данных

В работах [4, 5] предлагается разделять понятия качества данных и качества информации. Качество информации связано с внешней (потребительской) стороной и в этом случае под качеством понимается совокупность свойств, отражающих степень пригодности конкретной информации об объектах и их взаимосвязях для достижения целей, стоящих перед пользователем. Качество данных отражает внутренние вопросы, связанные с данными, их характеристиками, а так же с возможностями информационных систем по их поддержанию.

Качество информации и данных в информационных системах может быть рассмотрено и с точки зрения собственно данных (и информации) и с точки зрения обеспечения качества со стороны информационных систем. Это

позволяет разделить зону ответственности за качество между пользователями информационной системы и ее разработчиками.

Качество информации и данных определяется в терминах характеристик, описывающих внутренний и внешний взгляд на проблему качества. С внутренней точки зрения выделяют полноту, непротиворечивость, достоверность и корректность [4], и здесь в большей степени речь идет о качестве данных. С потребительской точки зрения интересует своевременность, релевантность, доступность, полезность и т.п. [5], и здесь в большинстве случаев речь идет о качестве информации. Некоторые характеристики качества могут быть рассмотрены и с внутренней и с внешней точек зрения. В работе [6] приводится суммарный отчет по множеству характеристик качества. В общем случае их выделяется 24, хотя многие из них взаимозависимы.

На основе характеристик качества можно сформулировать критерии качества данных. Критерии качества должны определяться не только разработчиками, но и пользователями КИС. Кроме того, изменения в сложных системах может приводить к серьезным проблем в задачах обеспечения качества, если только они не будут решаться некоторым стандартизованным для КИС образом совместными усилиями разработчиков и пользователей КИС.

Для решения этой проблемы предлагается использовать обобщенный репозиторий метаданных (ОРМД), описанный в работе [9]. Приведем здесь только основные элементы модели репозитория, которые интересны с точки зрения обеспечения качества данных и информации.

2. Обобщенный репозиторий метаданных

КИС можно представить как *совокупность автоматизированных бизнес-процессов, оперирующих понятиями предметной области*. К понятиям предметной области относятся, например, *Сотрудник, Организация*,

Подразделение, *Студент*, *Дисциплина*, и т.п. Понятия ИТ-области представляют собой *Сервер, Базу данных, Пользователя, Проект* и т.п.

Для описания понятий в объектно-ориентированном подходе [10] и в подходе, основанном на онтологиях [11], могут использоваться сущности. Сущность включает атрибуты и другие сущности. Между сущностями поддерживаются отношения агрегации, когда сущность включает другие сущности, ассоциации, когда сущности ассоциируются друг с другом, наследования, когда одна сущность содержит по крайне мере те же атрибуты и сущности, что и другая сущность, но может отличаться поведением.

В ОРМД сущности в большинстве случаев имеют проекцию на область базы данных, т.е. сущность связана с некоторой таблицей или представлением. Одна сущность обычно имеет одну базовую таблицу. Полное описание сущности может включать несколько таблиц. Атрибуты отображаются на переменные примитивных типов и обычно хранятся в базовой таблице сущности, являясь одним из полей этой таблице. Агрегированные сущности хранятся в таблице, не совпадающей с базовой.

Описание всех сущностей КИС хранится в ОРМД. Сущности отличаются друг от друга по имени, в рамках КИС имя сущности должно быть уникальным. Описание отношений между сущностями так же хранится в ОРМД. При описании отношений могут быть указаны не только типы отношений (агрегация, ассоциация или наследование), но и условия отношений. Например, *Пользователь КИС* может быть ассоциирован со *Студентом, Сотрудником, Внешним пользователем, Проектом* или *Серверной компонентной,* в зависимости от значений одного из атрибутов сущности *Пользователь КИС.*

Любого типа отношения могут быть контекстно-зависимыми и контекстно-независимыми. Отношения между сущностью *A* и сущностями *B* и *C* называются контекстно-зависимыми, если эти отношения определяются в зависимости от некоторого атрибута *D* сущности *A*. Примером таких отношений могут быть отношениями между сущностями *Пользователи КИС*

и *Студенты*, *Сотрудники и т.д.* Если между сущностями *A* и *B* существуют отношения, которые не зависят ни от одного атрибута *A* и *B*, то такие отношения называют контекстно-независимыми. Примером таких отношений может быть агрегация между *институтами* и *кафедрами*.

Рассмотрим теперь характеристики качества в КИС вуза и пути его повышения.

3. Полнота данных и информации

Полнота данных связана с представлением всех состояний систем реального мира в информационных системах [4, 7]. Рассмотрим на простом примере паспортных данных сотрудников вуза, чем отличается полнота данных от поддержки полноты данных со стороны КИС. Если информационная система содержит паспортные данные по всем сотрудникам вуза на некоторый момент времени, то можно говорить о том, что паспортные данные на этот момент времени являются полными. Если информационная система обеспечивает поддержку ввода паспортных данных не только граждан Российской Федерации (для которых обязателен ввод серии и номера), но и любых других, то можно говорить о поддержке полноты данных на уровне информационной системы. Таким образом, одна и та же область данных может быть полной на уровне данных и не полной на уровне поддержки информационной системы, а так же и наоборот.

Рассмотрим полноту данных на примере организации учебного процесса. Для определения критериев полноты ведения данных по организации учебного процесса рассмотрим некоторые понятия предметной области.

Институт – это подразделение университета, имеющее признак Института и занимающееся организацией образовательных программ. *Кафедра* – это подразделение университета, имеющее признак Кафедра и занимающееся реализацией программ и преподаванием дисциплин. *Образовательная программа* определяет цель учебного процесса, который организует *институт* и реализует *кафедра*.

Формальным критерием полноты данных может служить тот факт, что любой институт должен быть организатором хотя бы одной образовательной программы, а кафедра должна реализовывать по крайне мере одну образовательную программу или вести хотя бы одну дисциплину. Конечно, данный критерий не позволяет оценить полноту данных, так как в большинство случаев институт имеет не менее десятка образовательных программ. Этот критерии скорее является критерием «не пустоты» данных, тем не менее, мы его рассматриваем как минимальный критерий полноты.

Для того, чтобы реализовать описания основных критериев полноты данных, необходимо ввести функцию «Для каждого экземпляра сущности *A* существует хотя бы один ассоциированный с ним или агрегированный в него экземпляр сущности *B*». Далее эту функцию будем называть функцией проверки качества данных.

Примером сущности *A* может служить *Институт*, а сущности *B* – *Образовательная программа*. При задании в функции понятия *A* (или *B*), можно ограничить выбор каким-то условием по атрибутам понятия *A* (или *B* соответственно). Например, выбирать не все институты, а только те, которые принадлежат основному вузу, исключая филиалы, и только основные образовательные программы (для этого у сущности *институт* должен быть атрибут принадлежности к головному вузу, а у сущности *образовательная программа* должен быть атрибут тип программы со значением основная/дополнительная).

Для того чтобы определить отношения между экземплярами сущностей, для каждой пары сущностей строится матрица

$$E^{AB} = \{e_{ij}\}_{i=1,N}^{j=1,M},\ e_{ij} = \begin{cases} 1, & A_i -> C = B_j \\ 0, & A_i -> C \neq B_j \end{cases}., \qquad (1)$$

где *C* – атрибут сущности *A*, через который реализуется ассоциация, агрегация или наследование с сущностью *B*, *N*- число экземпляров сущности *A*, *M*- число экземпляров сущности *B*. Элемент e_{ij} определяет связь между *i*-

ым экземпляром сущности *A* и *j*-ым экземпляром сущности *B*. Тогда интересующим нас критерием будет:

$$\text{для}\ \forall i\ \exists j\ \text{такое что}\ e_{ij}=1,\ \text{а так же}\ \ \text{для}\ \forall j\ \exists i\ \text{такое что}\ \ e_{ij}=1. \qquad (2)$$

Это означает, что каждый институт должен организовывать по крайне мере одну образовательную программу, и каждая программа должна быть организована по крайне мере одним институтом.

В некоторых случаях критерий (2) будет содержать только первую или вторую часть. Например, учебные группы должны быть ассоциированы по крайне мере с одним студентом, но студенты не обязательно должны ассоциироваться с группой (студент может обучаться по индивидуальному плану).

Функция проверки качества в общем виде имеет две основные формы:

1. для любого экземпляра сущности *A* существует по крайне мере один ассоциированный с ним или агрегированный в него экземпляр сущности *B*;

2. объединение всех заданных условий выборки экземпляров сущности выделяет все экземпляры этой сущности.

Обсуждение второй формы функции проверки качества данных, связанной с определением условий для контекстно-зависимых отношений между сущностями и для бизнес-правил в определении бизнес-процесса можно найти в [9]. Вторая форма может быть описана через первую, если принять, что условие так же является сущностью: для любого экземпляра сущности *A* существует по крайне мере один экземпляр сущности *условие*, который с ним ассоциируется.

Процедуры верификации полноты данных в КИС строят матрицу (1) на основе данных из корпоративной базы данных и проверяют выполнение критериев (2), описанных в ОРМД. Процедуры проверки полноты данных не содержат блоки корректировки данных, но имеют блок информирования по электронной почте пользователей КИС о нарушениях критериев полноты.

Полнота информации связана с возможностью предоставления пользователям всей необходимой им информации для управления, анализа и

принятия решений. Для обеспечения полноты информации, во-первых, требуется, чтобы КИС покрывала все необходимые направления деятельности вуза. Во-вторых, для КИС необходима развитая система подготовки отчетов по всем направлениям деятельности, при этом требуется, чтобы отчеты связывали показатели различных направлений деятельности вуза. Отсюда, следует необходимость обеспечения интеграции корпоративных данных для обеспечения полноты информации. Интеграция корпоративных данных осуществляется с использованием репликаций, хранилищ и логической интеграции по требованию. Все формы интеграции описываются в ОРМД и в дальнейшем используются для построения отчетов пользователями КИС.

4. Достоверность данных и информации

Словарь определяет достоверность информации как свойство информации быть правильно воспринятой. Достоверность данных подразумевает, что все состояния информационной системы отображаются в состояние системы реального мира [4].

Достоверность данных часто связана с характеристикой – время, а, следовательно, и с актуальностью. На достоверность оказывает влияние то, насколько быстро выполнились процедуры актуализации, и восстановилось актуальное отражение данных реального мира в КИС.

Проблема актуальности данных в КИС возникает так же там, где речь идет о нескольких информационных системах, интегрированных между собой по крайне мере на уровне данных. Данные в информационной системе *B*, содержащей реплицированные данные информационной системы *A*, неактуальны, если выполняется одно из следующих условий:

1. информационная система *A* имеет данные, которые отсутствуют в *B*;
2. информационная система *B* имеет данные из *A*, которых на данный момент нет в *A*;

3. информационная система *B* содержит данные из *A*, которые в данный момент имеют другое состояние в *A*.

В рамках одной системы актуальность внутри КИС поддерживается разработчиками на уровне базы данных и на уровне приложения. Вопрос значительно усложняется, когда речь идет о логической интеграции данных, расположенных в различных базах данных. Актуализация осуществляется через триггеры, процедуры, службы и сервисы. В некоторых случаях актуализация выполняется в режиме реального времени (в этих случаях часто используются триггеры), для другого рода актуализации достаточно выполнение процедур, сервисов и служб один раз в час/день/неделю/месяц и т.п.

При выполнении актуализации следует учитывать все возможные изменения, связанные с рассогласованием данных. Например, в основных данных (справочниках) удалена запись, с которой в данных не был связан ни один экземпляр сущности, но которая использовалась в базах данных филиала вуза для одного или нескольких экземпляров.

Сценарий действия при возникновении этих проблем может быть прописан в коде программы, которая выполняет актуализацию, в коде хранимой процедуры или описан в ОРМД. Последнее является предпочтительным, так как позволяет администраторам всегда видеть всю схему актуализации данных, изменять/дополнять ее без изменения кода программы.

Сценарий действия в результате удаления элемента справочника предполагает замену значения атрибута на null или на другой элемент справочника. При этом сценарий должен включать не простое сопоставление элементов одного справочника между собой, а вычисляемую переменную, определяемую на основании, возможно, нескольких других атрибутов.

При формировании сценариев актуализации необходимо задавать условия и действия. Сценарий является связкой «если - то». При задании условий «если» используются сущности и их атрибуты. Рассмотрим

сущности *студент*, *сотрудник* и *пользователь КИС*. При удалении студента ассоциированный с ним пользователь КИС так же должен быть удален, если только он не является дополнительно и сотрудником. В последнем случае необходимо удалить только категорию студента у пользователя и связанные с ним права.

В ОРМД определен критерий достоверности данных: «для каждого *пользователя КИС* существует ассоциированный с ним *студен*т или *сотрудник*». Этот критерий строится на базе функции проверки качества данных, в которой разрешено объединение по «ИЛИ» экземпляров сущностей. Сценарий актуализации выглядит следующим образом: если для некоторого экземпляра сущности *пользователь КИС* нарушается критерий, то такой экземпляр должен быть удален.

При объединении по «ИЛИ» экземпляров сущности матрица (1) будет выглядеть следующий образом

$$E^{AB} = \left\{e_{ij}\right\}_{i=1,N}^{j=1,\sum_{k=1}^{K} M_k} \quad e_{ij} = \begin{cases} 1, & A_i -> C = B_{j_k}^{(k)} \\ 0, & A_i -> C \neq B_{j_k}^{(k)} \end{cases}, j = \sum_{l=1}^{k-1} M_l + j_k \ , \qquad (3)$$

где M_k - число экземпляров сущности $B^{(k)}$.

В некоторых случаях критерий (2) для элементов матрицы (3) для оценки достоверности данных может иметь ограничения на единственность: для $\forall i \ \exists! \, j$ такое что $e_{ij} = 1$. Например, любой *Студент* ассоциируется с единственным *пользователем КИС*.

Процедуры актуализации в КИС включают не только внесение/изменение/удаление записей в таблицы, но и выполнение более сложных сценариев. В управлении вузом есть несколько бизнес-процессов, которые должны автоматически поддерживаться процедурами актуализации данных:

- прием (увольнение), изменение должности сотрудника, перевод его в другое подразделение (или дополнительная работа в другом подразделении);

- зачисление (отчисление) студента, его перевод на другую специальность (другой факультет, курс, группу и т.п.);
- изменение организационной структуры вуза (слияние/перенос/удаление/создание/переподчинение/переименование подразделений).

Для примера рассмотрим бизнес-процесс отчисления студента. Сущность *Учебные планы студента* описывает отношение сущности *студент* к сущности *учебный план.*

Сценарий включает проверку критерия «для каждого экземпляра сущности *Студент* существует хотя бы один ассоциированный с ним экземпляр сущности *Учебные планы студента* с атрибутом статус = «Находится в процессе обучения». При нарушении данного критерия в сценарий включаются описание процедур, выполняющих обработку отчисления студента (в частности, *атрибут* экземпляра сущности *пользователь КИС* меняется на *бывший студент*, *пользователь КИС* лишается доступа к некоторым информационным ресурсам вуза и его личные каталоги на файловых серверах удаляются). В настоящее время сложные части сценария актуализации в ОРМД включают описание вызова процедур. Тем не менее, большая часть работы может быть описана в терминах сущностей, атрибутов и действий над ними.

5. Корректность данных

В работе [4] под корректностью понимается отсутствие искажений, когда система реального мира отображается в корректное состояние информационной системы. Под некорректным состоянием здесь понимается такое состояние, из которого нельзя вернуться ни к какому состоянию реальной системы или такое состояние, из которого можно вернуться в другое состояние системы реального мира.

Корректность данных чаще всего связана с вводом данных – ручным или автоматическим. Автоматический ввод обеспечивает меньшее число ошибок,

но, во-первых, на многих системах КИС невозможно организовать автоматический ввод, во-вторых, и при автоматическом вводе имеют место проблемы, связанные с качеством оборудования, предназначенного для ввода, и с качеством данных реального мира.

Большая часть данных в КИС вуза вносится вручную, и для обеспечения корректного ввода, необходимы как организационно-учебные мероприятия с участием персонала, ответственного за ввод, так и определенные механизмы на уровне КИС. Корректность данных может быть оценена, во-первых, с помощью специализированных проверок при вводе данных, во-вторых, путем автоматизированных процедур сопоставления данных при формировании отчетов, в-третьих, с помощью пользователей среды, которые получают одни и те же данные в различных приложениях, что повышает вероятность выявления ошибок ввода. Чем больше пользователей, сервисов среды и выше интенсивность работы, тем больше вероятность выявления некорректных данных и их исправления.

Определим первое правило ввода корректных данных: при вводе данных должны в максимально возможной степени задействоваться справочники. Использование справочников позволяет избежать многочисленных ошибок ручного ввода.

При вводе данных с помощью справочников остается проблема выбора некорректного элемента справочника. Частичным решением это проблемы является установление связей между справочниками. Справочники фактически являются реализацией некоторой сущности, и установление связей между справочниками осуществляется с помощью установления отношений между сущностями.

Помимо отношений между сущностями в ОРМД для справочников определены семантические связи. Например, справочник уровней образования (ВПО/СПО/НПО/СОО и т.п.) семантически связан со справочником уровней квалификаций (Магистр/Бакалавр/Специалист/Рабочий и т.п.). Эта связь выражается в том,

что при выборе высшего образования (ВПО) могут быть выбраны магистр, бакалавр или специалист, но не рабочий. В то же время, выбор специалиста позволяет выбирать ВПО или СПО, но не СОО.

Семантические связи между сущностями в ОРМД описываются специалистам предметной области. Это несколько удлиняет предварительный ввод данных, но дает значительные преимущества на момент ввода. Во-первых, уменьшается вероятность ввода некорректных данных, во-вторых, повышается удобство ввода данных, так как за счет семантических связей пользователь может выбирать альтернативы из небольшого объема данных.

В общем случае пусть имеется набор справочников $\{A_i\}_{i=1}^{N}$, каждый справочник состоит из набора элементов $\left\{a_i^j, i=\overline{1,N}, j=\overline{1,M_i}\right\}$. Здесь N - число справочников, а M_i - число элементов в i-ом справочнике. Для каждого элемента может быть определена связь с любым другим элементом любого справочника. $a_i^j \leftrightarrow a_k^l, 1 \le i,k \le N, 1 \le j \le M_i, 1 \le l \le M_k$.

Рассмотрим пример из трех справочников A_1, A_2, A_3 (Рисунок 1). Пусть в ОРМД описаны связи между $a_1^1 \leftrightarrow a_2^1$, $a_1^1 \leftrightarrow a_2^2$, $a_1^1 \leftrightarrow a_2^3$, $a_2^1 \leftrightarrow a_3^1$, $a_2^2 \leftrightarrow a_3^1$, $a_2^3 \leftrightarrow a_3^2$, $a_1^2 \leftrightarrow a_2^4$, $a_2^4 \leftrightarrow a_3^3$ и $a_1^2 \leftrightarrow a_3^2$.

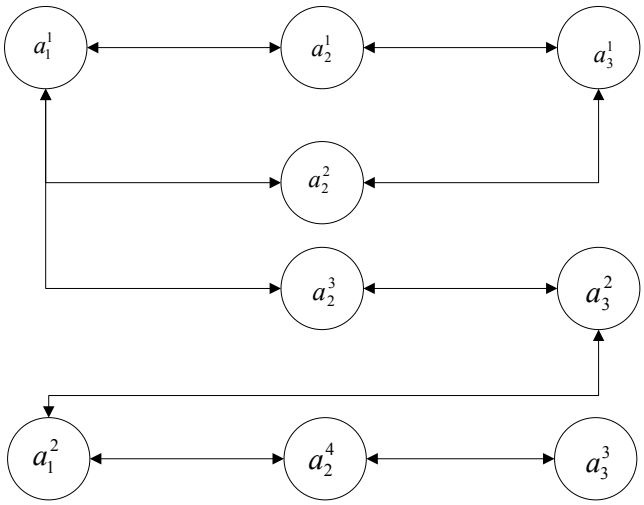

Рисунок 1. Пример семантических связей между справочниками

Рассмотрим, как происходит выбор элементов справочников. Пусть первоначально не задан ни один элемент из справочника. Выбор в справочнике A_1 элемента a_1^1 приведет к ограничениям в выборе в справочнике A_2 элементов a_2^1, a_2^2, a_2^3. Так как связь между A_1 и A_3 не определена, то при не заданном элементе из справочника A_2 допускается выбор любого элемента из справочника A_3, в том числе и элемента a_3^3, который не должен быть совмещен с a_1^1.

Для решения этой проблемы в ОРМД хранятся матрицы, описывающие связи (Таблица 1)

$$B = \left\{ b_{nm} = a_i^j \leftrightarrow a_k^l, n = \sum_{t=1}^{i-1} M_t + j, m = \sum_{t=1}^{k-1} M_t + l \right\}_{i,k=1,M}^{j=1,M_j, l=1,M_k}$$

$b_{nm} = 1$, если связь между элементами существует, иначе $b_{nm} = 0$.

Таблица 1.

Матрица связей элементов трех справочников

	a_1^1	a_1^2	a_2^1	a_2^2	a_2^3	a_2^4	a_3^1	a_3^2	a_3^3
a_1^1	1	0	1	1	1	0	0	0	0
a_1^2	0	1	0	0	0	1	0	1	0
a_2^1	1	0	1	0	0	0	1	0	0
a_2^2	1	0	0	1	0	0	1	0	0
a_2^3	1	0	0	0	1	0	0	1	0
a_2^4	0	1	0	0	0	1	0	0	1
a_3^1	0	0	1	1	0	0	1	0	0
a_3^2	0	1	0	0	1	0	0	1	0
a_3^3	0	1	0	0	0	1	0	0	1

Чтобы определить связи высокого порядка, т.е. связи которые описывают допустимые значения не напрямую, а через другие связи, необходимо

вычислить произведение матриц, в котором операции умножения и сложения заменены соответствующими логическими операциями:

$$a \bullet b = \begin{cases} 1, a = b = 1 \\ 0, a = 0 \mid b = 0 \end{cases} \quad a + b = \begin{cases} 0, a = b = 0 \\ 1, a = 1 \mid b = 1 \end{cases} \tag{4}$$

Кроме этого, при вычислении результата умножения матриц учитывается следующее

$$B^2 = B \bullet B = \left\{ b_{ij}^{(2)} = \sum_{k=1}^{\sum_{l=1}^{M} M_l} b_{ik} \bullet b_{kj}, i, j = \overline{1, \sum_{t=1}^{M} M_t} \right\}, \text{ при этом}$$

$$b_{ij}^{(2)} = a_n^m \left(\leftrightarrow\right)^2 a_t^l = \begin{cases} 0, n = t \ \& \ m \neq k \\ b_{ij}^{(2)}, n \neq t \mid (n = t \ \& \ m = k) \end{cases} \tag{5}$$

Выражение (5) означает, что связь между элементами одного и того же справочника не устанавливается (т.е. равна 0), а связь элемента с самим собой всегда равна 1.

Полная связь всех элементов справочника получается умножением матрицы B.

$$B^H = \prod_{i=1}^{\sum_{l=1}^{M} M_l} B \tag{6}$$

В выражении (6) следует учитывать правило (5), а так же операции логического умножения и сложения (4). Таким образом, при формировании допустимых значений для выбора элементов используется выражение (6), чтобы ограничить разрешенные элементы справочников.

Режим редактирования с использованием семантически связанных справочников может быть реализован двумя способами. Первый способ предполагает свободный выбор любого элемента из всех справочников без учета семантических связей. Проверка ввода выполняется при сохранении отредактированной информации на основании семантических связей. Второй способ предполагает автоматическую подстановку одного из возможных вариантов из семантически связанных справочников. В этой ситуации выбор некорректных данных невозможен.

Второе правило для обеспечения корректности данных – это использование процедур проверки ограничений на параметры ввода. Например, любые даты должны быть в оговоренных диапазонах (в том числе дата начала периода не должна быть позже даты окончания). Большая часть ограничений может быть реализована средствами системы управления базами данных (СУБД) – с помощью ограничений, триггеров и правила. Но возможности СУБД не всегда покрывают все требования.

Например, правилами вуза установлено, что абитуриент не может подавать заявление на более чем 3 программы высшего очного образования в один год. Такое ограничение может быть реализовано в коде программы (с возможностью менять только число доступных образовательных программ), в правилах или в триггерах (возможно в процедурах, которые вызываются из триггеров), или в полях базы данных. Если разработчики программы знают о возможности такого ограничения на момент разработки, то последнее решение является предпочтительным. Но, к сожалению, в процессе эксплуатации системы такие ограничения возникают там, где они ранее не были предусмотрены.

С точки зрения возможностей сопровождения КИС без участия программистов значительно более предпочтительно описать такие ограничения в ОРМД. Реализация такого подхода позволяет менять ограничения не только по числу образовательных программ, но и по другим параметрам в любое время без привлечения программистов.

Для того чтобы описать такое ограничение используются сущности и их атрибуты. Сущность *Учебная программа* имеет атрибутами *уровень учебной программы* и *год начала обучения*. *Учебная программа студента* имеет атрибутами ссылку на *учебную программу*, *студента* и *статус студента на учебной программе*.

Для того чтобы определить ограничение, нам необходимо в ОРМД ввести функции – количество, которая позволяет определять число экземпляров некоторой сущности. В ограничении мы должны указать, что число

экземпляров сущности *Учебная программа студента* с атрибутом статус = «Подано заявление» и заданным студентом и годом, при ограничениях на атрибут *уровень учебной программы* как «Высшее профессиональное образование» не должно превышать 3.

Если такое ограничение описано, то перед любым вводом или изменением в экземплярах сущности *Учебная программа студента* выполняется проверка ограничения, и оно не завершается, если ограничение нарушено.

Кроме счетчика экземпляров сущности в ограничениях могут использоваться операции сравнения, функции минимума и максимума.

Наконец, третье правило поддержки корректности данных связано с использованием уже введенных, возможно некорректных данных. Для повышения корректности данных следует их использовать различным образом:

1. опубликование данных в виде отчетов;
2. выполнение процедур на основе первичных данных, которые оказывают управляющее воздействие на системы реального мира;
3. формирование на основе первичных данных вторичной информации;
4. обработка данных для представления и принятия на их основе управленческих решений.

Опубликование данных часто связно с представлением некоторых отчетов для внутреннего или публичного доступа. Этот пункт позволяет пользователям самим контролировать корректность данных.

Примером управляющей процедуры может служить система единой регистрации пользователей КИС Владивостокского государственного университета экономики и сервиса (ВГУЭС) [12]. Все студенты ВГУЭС в первые дни учебы проходят регистрацию для создания учетной записи пользователя КИС. При регистрации предлагается ввести данные о ФИО и паспортные данные пользователя. Если введенные данные отличаются от тех, которые ранее внесены в базу данных в приемной комиссии, то студенты

обращаются в службы, которые занимаются ведением данных о студентах, с целью скорректировать паспортные данные или ФИО. Без регистрации студенты ВГУЭС не могут полноценно участвовать в учебном процессе и поселиться в общежитие, и, следовательно, ошибки ввода, связанные с ФИО и паспортными данными, будут выявлены почти в 100% случаях. Таким образом, процедуры управляющих воздействий позволяют повысить корректность данных.

Еще одна проблема связана с корректным формированием вторичных данных. Например, когда из студентов формируются группы, то возможно занесение студента не в ту группу. Проблемы с вторичными данными могут решаться несколькими способами. Первый связан с тем, что вторичная информация формируется через документы, результатом которых являются приказы. Сформированный приказ проходит несколько проверяющих инстанций, и после утверждения учебные группы сформируются автоматически, на основании подписанной электронной версии приказа.

Второй способ состоит в использовании вторичных данных пользователями КИС. Чем большее число пользователей использует вторичные данные и чем большая степень интеграции между вторичными и первичными данными из разных систем, тем большее число некорректных данных будет обнаружено.

Еще одна проблема связана с тем, что одна и та же ситуация в реальном мире может быть отображена на разные состояния информационной системы. В общем случае это не запрещено и даже более того, в этом случае информационная система работает корректно [4]. Но для получения агрегированных отчетов, необходимо иметь однозначное соответствие между состоянием системы реального мира и информационной системы.

Например, один и тот же работодатель может быть внесен в справочник более одного раза с разными названиями. При составлении агрегированных отчетов, например, получения числа выпускников вуза, проходивших

практику на предприятии, мы получаем некорректные данные, хотя первичные данные являются вполне достоверными и даже корректными.

Решение этой проблемы известно и описано в работах [6, 8], где предлагается вести таблицы соответствия элементов справочников (словарь синонимов) и построение агрегированных отчетов выполнять с учетом такого соответствия.

Непротиворечивость

Непротиворечивость согласно [4] означает, что не существует двух состояний системы реального мира, которые бы отразились в одно состояние информационной системы. Это позволяет однозначно восстановить из состояния информационной системы состояние системы реального мира.

Непротиворечивость данных связана, во-первых, с требованием обеспечить в информационных системах однозначное представление системы реального мира [5], во-вторых, с требованием получать одинаковые данные по одной и той же сущности при обращении в любую информационную систему КИС.

Определим понятие непротиворечивости данных для гетерогенной КИС. Непротиворечивость данных внутри КИС означает, что один экземпляр сущности должен иметь в КИС не более одного состояния по крайне мере в определенные в спецификации периоды.

Непротиворечивость информации по различным информационным системам в рамках КИС обеспечивается некоторым набором правил: первичный ввод данных в КИС осуществляется только в одном приложении; первичные данные могут храниться только на своем первичном сервере, откуда при необходимости они могут реплицироваться в другие базы данных. Внесение изменений в данные возможно лишь на первичном сервере. Следствием этого правила является необходимость иметь единые справочники для всех приложений КИС. Поскольку создание справочников

присутствует практически во всех проектах КИС, то можно выделить эту функциональность в отдельную подсистему создания справочников.

Единая система справочников не запрещает иметь отдельные справочники каждому проекту КИС. Как и возможность их редактировать и использовать. Единая система справочников лишь определяет иметь единственный справочник для одной сущности во всей КИС и использовать его в любом проекте среды.

Несмотря на наличие правила единственного справочника в КИС не исключено наличие противоречивых данных. Рассмотрим один из таких примеров.

В КИС ВГУЭС существуют две учетные записи пользователей с одним именем: пользователь Active Directory (AD) и запись в таблице базы данных, используемая для внешнего портала ВГУЭС [12]. Эти две учетные записи при создании идентичны, поэтому требуется, чтобы и множества учетных записей AD и множества учетных записей пользователей сотрудников и студентов внешнего портала совпадало с точностью до имени пользователя. Но КИС не может запретить администратору контроллера домена, создать вручную пользователя AD. Эта ситуация влечет за собой противоречивые данные по двум показателям. Во-первых, существует учетная запись AD с именем, которое уже есть в базе данных и в AD (в другом домене), во-вторых, в будущем может быть попытка создания такого пользователя, что приведет к трудно обнаруживаемой ошибке.

Критерий непротиворечивости данных с использованием функции оценки качества выглядит следующим образом: «для каждого экземпляра сущности *Пользователь портала* существует единственный экземпляр сущности *учетная запись AD*» и «для каждого экземпляра сущности *учетная запись домена AD* существует единственный экземпляр сущности *пользователь портала*». Если критерий непротиворечивости нарушен, то программа проверки критериев информирует об этом администратора.

Проблему непротиворечивости условий для управления бизнес-процессами и описания контекстно-зависимых отношений мы рассматривали в работе [9].

Доступность информации

Доступность информации обычно относят к пользовательским характеристикам качества [5]. Но мы определим доступность как внутреннее понятие КИС, так как связываем его с доступностью данных для пользователей КИС, основанное на системе управления правами [12]. Рассмотрим понятие доступность данных с одной стороны как обеспечение доступности по крайне мере одного пользователя к любому экземпляру любой сущности в КИС с правами на изменения, с другой стороны как способность КИС простым способом обеспечивать такой доступ для любого пользователя или любого множества пользователей КИС.

Решение обеспечения в КИС простого механизма управления доступом любого набора пользователей обсуждается в работе [12]. Там же можно найти и процедуры оценки доступности данных по областям видимости для пользователей в КИС.

В КИС связи между сущностями могут быть связями второго и более порядков. Рассмотрим пример, связанный с доступом пользователей к информационным ресурсам вуза. КИС вуза можно представить как совокупность информационной инфраструктуры, корпоративных данных и информационных систем. Каждая информационная система имеет своих пользователей, наделенных некоторыми ролями. Информационные системы представляют собой набор приложений, работающих с серверными компонентами. Каждая информационная система может вызывать методы нескольких серверных компонент. В свою очередь серверные компоненты могут вызывать методы других серверных компонент или обращаться к базам данных. Таким образом, можно определить связь между пользователями и понятиями.

Пусть $U = \{u_i, i = \overline{1,I}\}$ пользователи КИС, $P = \{p_k, k = \overline{1,K}\}$ - информационные проекты КИС, $F = \{f_n^m, m = \overline{1,M}; n = \overline{1,N_m}\}$ - методы серверных компонент; $D = \{d_l^j, l = \overline{1,L}, j = \overline{1,J_l}\}$ - сущности предметных областей КИС. Между этими понятиями определены связи первого порядка, которые в той или иной форме записаны в ОРМД или в системе управления правами [12].

$E^{UP} = \{e_{ik}^{UP}\} = \begin{cases} 1, u_i -> c = p_k \\ 0, u_i -> c \neq p_k \end{cases}$ связь пользователей и проектов

$E^{PF} = \{e_{knm}^{PF}\} = \begin{cases} 1, p_k -> c = f_n^m \\ 0, p_k -> c \neq f_n^m \end{cases}$ связь проектов и методов серверных компонент

$E^{FD} = \{e_{nmlj}^{FD}\} = \begin{cases} 1, f_n^m -> c = d_l^j \\ 0, f_n^m -> c = d_l^j \end{cases}$ связь методов серверных компонент и сущностей (здесь мы для общности не определяем характер связи – изменения/просмотр/удаление)

$E^{FF} = \{e_{n_1 m_1 n_2 m_2}^{FF}\} = \begin{cases} 1, f_{n_1}^{m_1} -> c = f_{n_2}^{m_2} \\ 0, f_{n_1}^{m_1} -> c \neq f_{n_2}^{m_2} \end{cases}$ связь методов серверных компонент друг с другом.

Из связей первого порядка могут быть получены связи второго и большего порядка, которые описывают связь между всем методами серверных компонент. Связь между пользователями и методами серверных компонент определяется как результат умножения матриц

$$E^{UF} = E^{UP} \bullet E^{PF}.$$

При умножении используется логическое умножение и сложение (4). Так как серверные компоненты связаны друг с другом, то можно определить полные косвенные связи между методами серверных компонент как

$$\tilde{E}^{FF} = \prod_{i=1}^{\sum_{m=0}^{M-1} N_m} E^{FF}.$$

Определим доступ пользователей к экземплярам сущностей

$$E^{UD} = E^{UP} \bullet E^{PF} \bullet \tilde{E}^{FF} \bullet E^{FD}.$$

Критерием доступности экземпляра сущности будет:

$$\text{для } \forall j \; \exists i \;\; \text{такое, что } e_{ij}^{UD} = 1. \qquad (7)$$

Это означает, что ко всем экземплярам всех понятий в КИС существует доступ по крайне мере у одного пользователя. Если для некоторого *j* критерий (7) не выполняется, то существуют пользователи, доступ которым ко всем сущностям КИС закрыт.

Интерес представляет так же наличие сущностей, к экземплярам которых отсутствует доступ хотя бы из одного проекта, т.е. следующий критерий не выполняется

$$E^{PD} = E^{PF} \bullet \tilde{E}^{FF} \bullet E^{FD}; \text{ для } \forall j \; \exists i \text{ такое, что } e_{ij}^{PD} = 1. \qquad (8)$$

Невыполнение условия (8) означает, наличие сущностей, которые нигде не используются и, возможно, которые никому не нужны.

Заключение

КИС вуза, достигая определенного уровня своего развития, становится либо движущей силой, либо тормозом на пути развития управленческих и образовательных процессов в вузе. Во многом это определяется тем, насколько КИС является управляемой. Одна из характеристик управляемости КИС связана с обеспечением качественной информации и агрегацией качественных данных. На поддержку вопросов качества данных и информации в КИС ВГУЭС тратится в среднем около 20% ресурсов, отведенных на разработку и сопровождение КИС.

Дальнейшие работы по поддержанию качества информации и данных в КИС ВГУЭС будут связаны с развитием механизмов автоматической поддержки качества на базе ОРМД.

Список литературы

[1] Терехов А.Н., Кияев В.И., Комаров С.Н. Принципы информатизации системы управления в Санкт-Петербургском Государственном Университете//Вестник С-Пб. ун-та. сер. 8, 2004, вып.2 (№ 16), с. 187-201.

[2] Wand Y., Weber R. An ontological model of an Information System//IEEE Trans. Soft. Eng. 16 11. – 1990. - pp. 1282-1292.

[3] Science Magazine. Complex Systems.-1999. Vol. 284. #5411.pp 1-212.

[4] Wand Y., Wang, R. Anchoring Data Quality Dimensions in Ontological Foundations//Communications of the ACM. – 1996.- November.- pp. 86-95.

[5] Price R., Shanks G. A Semiotic Information Quality Framework//Proc. IFIP International Conference on Decision Support Systems (DSS2004): Decision Support in an Uncertain and Complex World, Prato.-2004.

[6] Wang R., Storey V., Firth C. A framework for analysis of data quality research//IEEE Trans. on Knowl. Data Eng. -1995.-7, 4. - pp.623-640.

[7] Wang R., Ziad M., Lee Y.W. Data Quality. Kluwer 2001. p.167.

[8] Madnick, S.; Wang, R.; and Xian, Xiang. The Design and Implementation of a Corporate Householding Knowledge Processor to Improve Data Quality//Journal of Management Information Systems. – 2004.-vol. 20,-№3.-pp.41-69.

[9] Шахгельдян К.И. Корпоративная информационная среда: подход, основанный на понятиях//Информационные технологии моделирования и управления.-2006.-№4 (29).-с.503-510.

[10] Буч Г., Рамбо Д., Джекобсон А. UML. Руководство пользователя. –М: ДМК, 2000. – 432с.

[11] Клещев А.С., Шалфеева Е.А. Классификация свойств онтологий. Онтологии и их классификации. Препринт, Владивосток, ИАПУ ДВО РАН, 2005, 19с.

[12] Шахгельдян К.И., Крюков В.В., Гмарь Д.В. Система автоматического управления правами доступа к информационным ресурсам вуза//Информационные технологии 2006.- №2 -с.19-29.

Shakhgeldyan K.J.

The challenges of data and information quality in University's Information Environment

Resume

The challenges of data and information quality are the main subjects of the article. The classical definitions of data quality dimensions are extended for complicate information environment. The methods providing data quality are discussed too.

сведения об авторах

Шахгельдян Карина Иосифовна, к.т.н., Владивостокский государственный университет экономики и сервиса, начальник отдела информационных сервисов и корпоративных приложений

Домашний адрес – Владивосток, ул. Пологая 62, кв. 22, тел.: (4232)-433252

Служебный адрес – Владивосток, ул. Гоголя 41, тел.: (4232)-404226

Паспортные данные 05 03 № 804365, выдан 22.12.2003 , Ленинским РУВД г. Владивостока

г.р. 1967.

Страховое св-во: 037-809-818-90