ТЕХНОЛОГИЯ ИНТЕГРАЦИИ ГЕТЕРОГЕННОГО КОНТЕНТА В КОРПОРАТИВНЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

***Зыков Сергей Викторович**, кандидат технических наук, доцент Национального исследовательского университета – Высшей школы экономики, Москва*
E-mail: szykov@hse.ru

В работе представлены основные положения новой технологии интеграции разнородных данных. В современных корпорациях накоплены значительные объемы данных, которые растут экспоненциально. Актуальность проблемы манипулирования такого рода данными усугубляется их гетерогенностью. Разработанная технология включает комплекс новых математических объектных моделей, методов и программных средств для представления и манипулирования данными.

***Ключевые слова:** интеграция данных, модель данных, управление контентом.*

HETEROGENEOUS CONTENT INTEGRATION TECHNOLOGY FOR ENTERPRISE INFORMATION SYSTEMS

***Sergey Zykov**, Ph.D., Associate Professor, National Research University – Higher School of Economics, Moscow*
E-mail: szykov@hse.ru

The paper presents major concepts of the new technology for large–scale data integration. The current enterprise data size is large, and it grows exponentially. The data management issue is even more challenging due to the heterogeneous character of the data. The technology developed encompasses a set of new object–based models, methods and software tools for representing and manipulating heterogeneous data.

***Keywords:** data integration, data model, content management.*

Введение

В настоящее время в корпоративных структурах уже накоплены и лавинообразно нарастают весьма значительные объемы разнородных данных, которые в ряде случаев достигают петабайт. Попытки построения интегрированных схем разнородных данных средствами так называемых «промышленных» методологий проектирования (в том числе IBM RUP, Microsoft MSF, Oracle CDM), не поддержанные теоретическими обобщениями объектного типа, приводят либо к неоправданно узкому спектру решений от единственного поставщика, либо к неадекватным затратам по времени и стоимости. С другой стороны, существующие теоретические подходы к моделированию и интеграции информационных систем (такие как категории [1], онтологии [2], проект «СИНТЕЗ» [3]) в силу отдаленности от современных промышленных технологий (в том числе инструментальных средств поддержки разработки – CASE, быстрого построения прототипов – RAD) не приводят к созданию программных комплексов с практически приемлемыми эксплуатационными характеристиками, включая масштабируемость, эргономичность и расширяемость. Актуальность отмеченных проблем также подтверждается значительными ассигнованиями по перечисленным направлениям целого ряда федеральных и международных программ в РФ, ЕС, США, ООН, и ЮНЕСКО.

В этой связи разработан подход к интеграции данных в гетерогенных программных комплексов, взаимодействующих в глобально распределенной среде, основанный на математических моделях и поддержанный инструментальными средствами, обеспечивающими сопряжение со стандартными CASE-средствами для «промышленных» методологий построения ПО. В основу подхода положена многоуровневая технологическая схема для интеграции данных в корпоративных программных комплексах (КПК), более подробно описанная далее. Указанная технологическая схема для интеграции данных поддерживается системой

моделей для представления и манипулирования объектами данных в информационных системах в составе КПК. При этом система моделей реализована по каскадному принципу, так что выход модели для представления интегрированной схемы объектов данных – в форме описания класса на языке UML – является входом модели для манипулирования объектами данных. Каждый из перечисленных видов моделей поддержан предметно-ориентированным визуальным инструментальным средством. Модель для представления интегрированной схемы объектов данных поддержана программным средством для семантической интеграции данных в информационных системах в составе КПК под названием ConceptModeller. Модель для манипулирования объектами данных поддержана программным средством для управления контентом под названием «Инструментальное средство управления контентом» или ИСУК. Инструментальные средства ConceptModeller и ИСУК, как и соответствующие им модели, связаны между собой по принципу каскада. Перечисленные программные средства вместе с информационными системами в составе КПК объединяются на базе унифицированной портальной архитектуры, которая позволяет строить на общей основе быстрые прототипы и полномасштабные реализации КПК с интеграцией данных для ряда корпоративных производственно-сбытовых структур. При этом применение указанных математических моделей и программных средств в комплексе с портальной архитектурой открывает возможность построения на общей основе портальных интерфейсов для корпоративных сотрудников (интранет), клиентов и партнеров (экстранет), а также для пользователей сети интернет в виде официального веб-сайта корпорации. Существенно, что сценарно-ролевые соотнесения позволяют не только реализовать удобный и интуитивно ясный интерфейс для каждой из названных категорий пользователей за счет персонализации, но и обеспечить для всех пользователей необходимый уровень безопасности данных.

Кроме того, следует отметить, что предложенный интеграционный подход обеспечивает устранение дублирования и противоречий на уровне данных в корпоративных программных комплексах, что существенно повышает надежность информационных систем в составе КПК. При построении технологии учтен целый ряд взаимосвязанных факторов, оказывающих влияние на интеграцию гетерогенных данных, представленных объектами, в том числе систем программирования, моделей данных, методологий и средств разработки программных систем, архитектур ИС и СУБД.

Модели интеграции гетерогенных данных в корпоративных информационных системах

Для обеспечения адекватного моделирования схемы интеграции гетерогенных данных в программных системах и комплексах корпоративного типа разработан комплексный подход, включающий объектные модели как для представления данных, так и для организации управления ими.

Общая технологическая схема интеграции данных в информационных системах в составе КПК обеспечивает замкнутый, двунаправленный характер их построения, а также потенциально открывает возможность реинжиниринга [4,7]. Последняя возможность весьма важна для осуществления верификации информационных систем в составе КПК на уровне данных, существенно повышающей надежность их функционирования.

Технологическая схема интеграции данных в информационных системах в составе КПК содержит этапы, соответствующие формам представления информации в гетерогенных программных комплексах, взаимодействующих в глобально распределенной среде (естественный язык, математические модели, сопряжение с инструментальными средствами, управление контентом и др.) и уровни, детализирующие эти этапы (объекты, связи, события, примеры инструментальных средств и программных систем).

Контент-ориентированный подход позволяет обобщить на базе объектных моделей понятия данных и метаданных, унифицировать управление обработкой гетерогенных объектов, а также адекватно моделировать интернет-среду, что важно для обеспечения надежности информационных систем в составе КПК.

Объектный характер моделей, формируемых по схеме «класс-объект-значение» позволяет обеспечить преемственность с традиционным объектно-ориентированным подходом к анализу и разработке ПО (Object-oriented analysis and development, OOAD), модели «сущность-атрибут-значение» (entity-attribute-value, EAV), а также теоретически перспективными подходами (КМ Вольфенгагена [5], переменных областей Д.Скотта [6]) и развить их в направлении интернет-среды [4,7-9].

Технологическая последовательность преобразований в модели в общем виде выглядит следующим образом. Прежде всего, строится терм конечной последовательности представляющий описание класса объекта данных информацион-

ной системы на формальном языке, например, l-исчислении. Далее происходит построение предиката, содержащего ограничения, накладываемые предметным аналитиком на данный класс объектов в рамках проектируемой интегрированной схемы данных информационной системы в составе КПК. При этом могут использоваться логические операции и причинно-следственные связки, а также кванторы применяемые в логике высших порядков. Затем описание класса трансформируется в эквивалентное графическое представление в форме фрейма. В действительности эксперт-аналитик, как правило, работает с интегрированными схемами данных уже на уровне этого графического представления, или, что в ряде случаев более предпочтительно, на уровне UML-диаграмм, визуализирующих XML-описания классов объектов данных в терминах программного средства ConceptModeller. Отмеченные уровни представлений данных реализованы в рамках созданной технологии и позволяют осуществить гибкий переход от концептуальной схемы данных на семантических сетях с фреймовой визуализацией к схеме данных, представленной в стандартном формате UML, обеспечивающем совместимость с современными инструментальными средствами разработки информационных систем и КПК уровня CASE.

При этом модель представления контента основана на ситуативной интерпретации в форме семантических сетей, что обеспечивает интуитивную ясность для предметных экспертов при построении схемы предметной области; удобство использования модели обеспечивается графическим представлением в виде фреймов. Модель организации управления контентом организована в виде абстрактной машины с состояниями и ролевыми соотнесениями, что позволяет естественно обобщить процессы, существующие в большинстве подобных инструментальных средств: построение шаблонов веб-страниц, редакторский цикл публикации веб-страниц, разграничение ролей, доступа и др. При этом важнейшие операции по управлению контентом – определение, означивание, персонализированное манипулирование и др. – формализованы посредством оригинального языка абстрактной машины, для которого разработаны формальный синтаксис и денотационная семантика в терминах переменных доменов (включая порядок построения объектов контента, семантические функции и предложения для этих операций) [4,8].

Технологическая последовательность преобразований в модели выглядит следующим образом. Прежде всего, формируется описание классов в форме термов теории вычислений Д.Скотта, включающих имена доменов и операции-конструкторы – декартово произведение, функциональное пространство, последовательность и дизъюнктная сумма. Затем происходит построение функции над доменами в терминах логики высших порядков, которое дополнительно вовлекает причинные (причина, следствие) и логические (И, ИЛИ, НЕ) операции. Далее, строится эквивалентное графическое представление функции над доменами в форме фрейма, т.е. отношения на концептах. Концепты фрейма изображаются вершинами (овалами), отношения – направленными дугами (стрелками), а операции – охватывающими операнды прямоугольниками. Каждый элемент фрейма может иметь имя. После этого происходит формирование XML-объекта с описанием класса в стандарте UML, эквивалентного определенному фрагменту фрейма с конкретизацией фиксированных элементов метаданных, что соответствует шаблону веб-страницы ИСУК. Наконец, конкретизация всех элементов данных и метаданных шаблона позволяет получить HTML-представление, соответствующее коду веб-страницы ИСУК на портале КПК.

Архитектурная схема объединенного хранилища гетерогенного корпоративного контента обеспечивает унификацию за счет использования обобщенных объектных ассоциативных связей на уровне гетерогенных данных и метаданных. С другой стороны, унификация манипулирования контентом гетерогенных информационных систем основана на использовании единой мета-надстройки над корпоративным хранилищем данных в форме Интернет-портала. При этом динамическое, сценарно-ориентированное управление контентом в рамках портальной архитектуры обеспечивается соотнесениями, реализованными в форме сценариев языка программирования, изменяющими состояния абстрактной машины. Абстрактная машина на состояниях, таким образом, выступает в качестве модели манипулирования корпоративным контентом.

Сценарии другого рода реализуют персонализированное манипулирование контентом, поддержанное математической моделью в форме многопараметрического функционала, а также инструментальным средством ИСУК.

Инструментальные средства интеграции данных и манипулирования контентом

Инструментальное средство ConceptModeller позволяет осуществить семантически ориентированное визуальное проектирование интегриро-

ванной схемы данных гетерогенного корпоративного программного комплекса. При этом применяется математическая модель с семантическими сетями, обеспечивающих работу в терминах, близких к естественно-языковым и ясных предметному эксперту. Визуализация основана на фреймовом представлении схемы данных.

Таким образом, за счет интеграции с разработанными математическими моделями и современными инструментальными средствами ConceptModeller обеспечивает замкнутый, непрерывный цикл проектирования – от математической модели до инструментальной схемы данных. Отметим возможность реинжиниринга, то есть расширения, развития и совершенствования КПК на стадии сопровождения как за счет так доработки существующих, так и посредством включения новых информационных систем, с сохранением при этом общей схемы интеграции на уровне данных. При этом описания фреймов в терминах инструментального средства ConceptModeller представляют собой упорядоченные списки вида: «атрибут-тип-значение».

Инструментальное средство ИСУК имеет в своей основе модель в форме абстрактной машины и позволяет осуществить предметно ориентированное визуальное манипулирование гетерогенным объектами данных (или контентом) для информационных систем в составе КПК, а также размещение контента на корпоративном портале. Особенностями ИСУК являются гибкость редакторского цикла и механизма ролей, обеспечивающих доступ к контенту на основе динамически корректируемых профилей доступа и шаблонов веб-страниц. Благодаря сценарно-ориентированному управлению контентом ИСУК обеспечивает унификацию представления гетерогенных объектов данных и метаданных на портале, гибкое взаимодействие различных классов пользователей (рядовых и привилегированных, корпоративных и внешних) с контентом, высокую надежность данных (на основе сценариев и профилей разграничения доступа), а также повышенный уровень эргономики (с учетом персональных предпочтений) и прозрачное манипулирование сложными объектами данных (в т.ч. мультимедиа). При этом классы представляют собой упорядоченные списки вида «атрибут-тип», а шаблоны – упорядоченные списки вида: «атрибут-тип-значение».

Заключение

Практическое применение предложенного подхода к интеграции данных для информационных систем в КПК позволило реализовать серию унифицированных корпоративных программных комплексов для различных отраслей промышленности, объединяющих гетерогенные компоненты, в том числе – модули современных систем на базе Oracle для финансового планирования и управления, унаследованную систему учета людских ресурсов, а также слабоструктурированные архивы мультимедиа-объектов, включая отсканированные документы, аудио- и видеоданные. Реализация серии интернет- и интранет-порталов, манипулирующих контентом гетерогенных корпоративных программных комплексов, обеспечила целый ряд внедрений в различных компаниях многопрофильной международной группы «ИТЕРА», объединяющей около 10 тысяч сотрудников в 150 компаниях 24 стран мира.

Комплексный характер технологии, которая включает математические модели, инструментальные средства, а также портальные архитектурно-интерфейсные решения, обеспечивает сопряжение с широким диапазоном современных инструментальных средств (IBM Rational, Microsoft Visual Studio .NET, Oracle Developer) и современными стандартами (UML, XML) разработки информационных систем и КПК.

Функциональные преимущества подхода по сравнению с выявленными аналогами – манипулирование сложными, разнородными и в разной степени структурированными объектами данных, интеграция на уровне данных компонент с различной архитектурой – обусловлены ориентацией моделей и инструментальных средств на гетерогенные портальные корпоративные программные комплексы. Качественные оценки функциональных возможностей подхода подтверждены сравнением важнейших макропоказателей – совокупной стоимости владения, возврата инвестиций и сроков внедрения. Результаты внедрения в МГК «ИТЕРА», с учетом существенной гетерогенности объектов данных корпоративных хранилищ, по перечисленным показателям превосходят передовые коммерческие аналоги в среднем на 30-40%.

Результаты исследования, а также созданные на их основе программные комплексы, учебные программы и курсы внедрены в целом ряде коммерческих и государственных структур корпоративного типа, прежде всего, Международной группы компаний «ИТЕРА», Института проблем управления РАН, Минпромэнерго РФ, а также холдингов ЛАНИТ, Softline и Luxoft [4,7,9].

Литература:

1. Вольфенгаген В.Э., Косиков С.В. Аппликативные системы для представления знаний.– в кн. Принципы построения и технология проектирования систем искусственного интеллекта.– Вып. 3, 1989, с. 17-19.
2. Клещев А.С., Артемьева И.Л. Математические модели онтологий предметных областей.– Ч.2. Компоненты модели. // НТИ, сер. 2, 2001, № 3.– с.19-29.
3. Калиниченко Л.А. СИНТЕЗ: язык определения, проектирования и программирования интероперабельных сред неоднородных информационных ресурсов.– М.: ИПИ РАН, 1993.
4. Зыков С.В. Основы проектирования корпоративных систем.– М.: НИУ ВШЭ, 2012. – 432 с.
5. Вольфенгаген В.Э., Яцук В.Я. Аппликативные вычислительные системы и концептуальный метод проектирования систем знаний. — М: МО СССР, 1987.
6. Скотт Д.С. Области в денотационной семантике.– с.58-118.– в кн.: Математическая логика в программировании.– М.: Мир, 1991. – 408 с.
7. Зыков С.В. Системная инженерия корпоративных программных комплексов // Материалы IV Научно-практической конференции «Актуальные проблемы системной и программной инженерии». – М.: ИД НИУ ВШЭ.– 2015.– с.48-54.
8. Zykov S.V. ITERA Enterprise Portal: from Model to Implementation. In: International Conference on Enterprise Information Systems and Web Technologies, ISRST: Orlando, FL, USA, July 13-16, 2009, pp.140-145.
9. Zykov S.V. Enterprise Content Management: Bridging the Academia and Industry Gap In: N.Callaos, W.Lesso, C.D.Zinn, B.Zmazek (Eds.), Proc. of International Conference on Information Society (i-Society 2007), Merrillville, Indiana, U.S.A., October 7-11, 2007, Vol. I, pp.145-152.

References:

1. Wolfengagen V.E., Kosikov S.V. Applikativnye sistemy dlya predstavleniya znanii. – v kn. Principy postroeniya i tehnologiya proektirovaniya sistem iskusstvennogo intellekta.– Vyp.3, 1989. – s.17-19 (Wolfengagen V.E., Kosikov S.V. Applicative Systems for Knowledge Representation.– In: Artificial Intelligence Systems: Construction Principles and Design Technology, Vol.3, 1989. – pp.17-19)
2. Kleshev A.S., Artemyeva I.L. Matematicheskiye modeli ontologii predmetnyh oblastei.– Ch.2. Komponenty modeli. // NTI, ser. 2., 2001, No.3.– c.19-29 (Kleshev A.S., Artemyeva I.L. Mathematical Models for Problem Domain Ontologies.– Part II.– Model Components // NTI, Vol.2., 2001, No.3.– pp.19-29).
3. Kalinichenko L.A. SINTEZ: yazyk opredeleniya, proektirovaniya i programmirovaniya interoperabelnyh sred neodnorodnyh informatsionnyh resursov.– M.: IPI RAN, 1993 (Kalinichenko L.A. SINTEZ: The Language to Define, Design and Implement Interoperable Environments for Heterogeneous Information Resources.– M.: IPI RAN, 1993).
4. Zykov S.V. Osnovy proektirovaniya korporativnyh sistem.– M.: NIU VSHE, 2012. – 432 s. (Zykov S.V. Enterprise Systems Design: Foundations.– M.: NIU VSHE, 2012. – 432 pp.)
5. Wolfengagen V.E., Yatsuk V.Ya. Applikativnye vychislitel'nye sistemy i kontseptual'nyi metod proektirovaniya sistem znanii.– M: MO SSSR, 1987. (Wolfengagen V.E., Yatsuk V.Ya. Applicative Computing Systems and Conceptual Method of Knowledge Systems Design.– M: MO SSSR, 1987.
6. Scott D.S. Oblasti v denotatsionnoi semantike.– s.58-118.– v kn.: Matematicheskaya logika v programmirovanii. – M.: Mir, 1991. – 408 s. (Scott D.S. Domains in Denotational Semantics.– pp.58-118.– In.: Mathematical Logics in Programming. – M.: Mir, 1991. – 408 pp.)
7. Zykov S.V. Sistemnaya inzheneria korporativnyh programmnyh kompleksov // Materialy IV Nauchno-prakticheskoi konferentsii "Aktual'nye problemy sistemnoi i programmnoi inzhenerii". – M.: ID NIU VSHE.– 2015.– s.48-54. (Zykov S.V. System Engineering for Enterprise Software Systems // Proceedings of 4th Conference on Problems of Systems and Software Engineering". – M.: ID NIU VSHE.– 2015.– pp.48-54.

Zykov S.V. ITERA Enterprise Portal: from Model to Implementation. In: International Conference on Enterprise Information Systems and Web Technologies, ISRST: Orlando, FL, USA, July 13-16, 2009, pp.140-145.

Zykov S.V. Enterprise Content Management: Bridging the Academia and Industry Gap In: N.Callaos, W.Lesso, C.D.Zinn, B.Zmazek (Eds.), Proc. of International Conference on Information Society (i-Society 2007), Merrillville, Indiana, U.S.A., October 7-11, 2007, Vol. I, pp.145-152.