

Модели данных для информационного взаимодействия

А.Б. Корчагин, И.Г. Лисьих, Д.А. Никифоров, Р.Л. Сиваков

Аннотация – Данной публикацией открывается цикл статей, посвященных анализу различных подходов к моделированию данных и выявлению их ключевых технологических особенностей. Рассмотрена зависимость уровня сложности структур данных от уровня организации информационного взаимодействия. Предложена структура паспорта информационной модели (набора учетных сведений) и приведены паспорта некоторых моделей данных: UN/CEFACT CCL, WCO Data Model, ISO 20022, NIEM, WIPO St.96.

Ключевые слова – информационное взаимодействие; модель данных; модель процессов; элемент данных; метамодель данных; документ; сообщение; совместимость моделей данных; отображение моделей данных; онтология.

I. ВВЕДЕНИЕ

Рост интереса к моделям данных в сфере информационного взаимодействия обусловлен развитием сервисов сети Интернет, необходимостью удовлетворения потребностей бизнес-сообщества и административных органов, расширением взаимодействия в электронной форме на межведомственном и межгосударственном уровнях. При этом существенно возрастает уровень сложности используемых информационных систем и, как следствие, повышаются требования к организации передаваемых данных, к их согласованности и непротиворечивости.

По характеру интересов и роду практической деятельности авторы много лет занимаются темой моделирования информационных систем, поэтому предполагают, что их опыт будет полезен в первую очередь бизнес- и системным аналитикам, системным архитекторам, директорам ИТ-служб, в целом лицам, принимающим решения при разработке информационных систем.

В заявленном цикле статей авторы не будут касаться темы моделирования собственно информационного взаимодействия и процессов в целом, если этого не требуется для понимания той или иной модели данных. Эта

Статья получена 09 февраля 2017.

А.Б. Корчагин – кандидат технических наук, доцент Уральского федерального университета, ведущий консультант компании Центр ИТ (e-mail: Alexander.Korchagin@centre-it.com).

И.Г. Лисьих – ведущий консультант компании Центр ИТ (e-mail: Igor.Lisyih@centre-it.com).

Д.А. Никифоров – бизнес-аналитик компании Центр ИТ (e-mail: Denis.Nikiforov@centre-it.com).

Р.Л. Сиваков – президент группы компаний «ЦИТ» (e-mail: Ruslan.Sivakov@centre-it.com).

тема не менее актуальна, чем моделирование данных, о чем свидетельствует наличие соответствующих методик и моделей, разработанных под патронажем авторитетных международных организаций и ИТ-вендоров, таких, как UN/CEFACT, APQC, APICS, IBM и др. Предполагается, что направление моделирования процессов будет темой самостоятельного исследования.

Прежде всего следует уточнить ряд вопросов, касающихся терминологии.

Термин «модель данных» у специалистов-практиков очень часто ассоциируется исключительно с базами данных. Это обусловлено тем, что он возник именно в контексте проектирования баз данных, его впервые формально определил основоположник теории реляционных баз данных Э. Кодд [1]. В дальнейшем термин получил широкое распространение, благодаря работам крупнейших специалистов в области баз данных, таких, как К. Дж. Дейт [2], Хью Дарвен [3] и др.

Между тем, термин «модель данных» все чаще используется в контексте проектирования структур данных, составляющих предмет обмена при взаимодействии информационных систем, в общем случае – гетерогенных. Именно такие модели данных рассматриваются в настоящей работе, при этом авторы будут придерживаться следующего определения, приведенного в стандарте ISO 11179: «модель данных (data model) – графическое и (или) словесное представление данных, задающее их структуру и взаимосвязи» [4].

Отметим, что в ряде описаний структур данных для самообозначения не всегда используется термин «модель данных» и даже просто «модель». Тем не менее, они, несомненно, являются моделями данных, если соответствуют определению из ISO 11179. Примером является Библиотека ключевых компонентов СЕФАКТ ООН (The UN/CEFACT Core Component Library, UN/CCL).

Другим важным понятием в области моделирования данных является «элемент данных». В различных источниках это понятие определяют по-разному. В контексте баз данных его используют достаточно редко [5] и обычно отождествляют с понятием «тип» [2]. В контексте структур данных понятие «элемент данных» обозначает структурную единицу данных, снабженную характеристиками, позволяющими отличить ее от других единиц данных. В стандарте ISO 11179 дано следующее определение: «элемент данных (ЭД) (data element, DE) – единица данных, для которой с помощью набора атрибутов заданы определение, идентификация, представление и допустимые значения» [4].

Для понятия «элемент данных» так же, как и для модели данных, в некоторых случаях используют альтер-

нативные термины. Так, в уже упомянутой Библиотеке ключевых компонентов СЕФАКТ ООН, используют термины «ключевой компонент» (Core Component) и «сущность бизнес-информации» (Business Information Entity).

II. УРОВНИ СЛОЖНОСТИ СТРУКТУР ДАННЫХ И УРОВНИ ОРГАНИЗАЦИИ ИНФОРМАЦИОННОГО ВЗАИМОДЕЙСТВИЯ

При анализе эволюции подходов и технических решений, применяемых в ходе проектировании и реализации информационного взаимодействия, прослеживается определенная зависимость уровня сложности структур данных от уровня организации (упорядоченности) информационного взаимодействия, что в схематическом виде представлено на рисунке 1. Отметим, что эта зависимость хорошо ложится на траекторию эволюционных изменений систем в пространстве координат по направлениям «простота – сложность» и «хаос – порядок» [6], считая, что уровень организации информационного взаимодействия лежит в направлении «хаос – порядок».

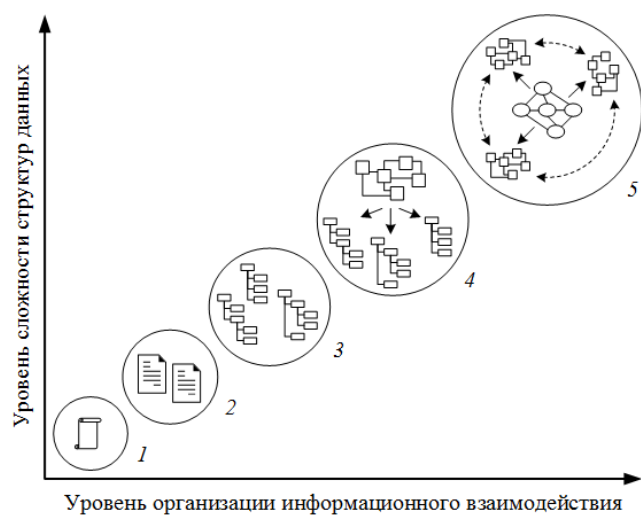


Рис. 1. Зависимость степени сложности структур данных от уровня организации информационного взаимодействия: 1 – неуправляемое взаимодействие; 2 – регламентированное взаимодействие; 3 – обособленное взаимодействие; 4 – внутриотраслевое взаимодействие; 5 – межотраслевое взаимодействие

Следует сказать, что, хотя здесь речь идет о качественных оценках уровней организации и сложности, можно рассматривать и вполне измеримые характеристики. Для сложности структур данных в качестве меры может служить количество взаимосвязей для элементов данных. Уровень организации взаимодействия можно оценивать по количеству охваченных бизнес-процессов, количеству процедур, количеству участников взаимодействия и др.

Неуправляемое взаимодействие соответствует обмену данными в неструктурированном виде (например, личное письмо, записка). Уровень сложности структур данных – «тривиальный». Здесь отсутствует какая-либо процессная регламентация и, соответственно, не выдвигаются каких-либо особых требований к данным, кроме низкоуровневых (например, ограниченный набор сим-

волов). Такой обмен, как правило, осуществляется на нерегулярной основе и обычно ассоциируется с бумажными носителями, но сохраняется и в электронной форме в виде электронной почты, Skype, SMS и др. Модели данных для тривиально уровня не имеют смысла.

Регламентированное взаимодействие подразумевает обмен данными с использованием установленных форм документов или бланков (внутрикорпоративные документы, налоговая декларация, формы статистической отчетности и др.). Уровень сложности структур данных – «формализованный». Такой обмен может осуществляться как на регулярной, так и нерегулярной основе и также ассоциируется с бумажными носителями. Даже если документы формируются и передаются с использованием средств вычислительной техники, какой-либо автоматизированной обработки содержащихся в них данных не предполагается. При информационном обмене в электронной форме данный уровень соответствует, например, передаче файлов в формате PDF или сканкопий бумажных документов средствами электронной почты. Некоторым аналогом модели данных для формализованного уровня является описание формы документа и правил ее заполнения в регламентирующем нормативно-правовом акте.

Обособленное взаимодействие предполагает обмен данными по цифровым каналам связи с использованием унифицированных форм документов в структурированном виде (налоговая декларация, формы статистической отчетности и др., разложенные на элементы данных). Уровень сложности структур данных – «структурированный». Такой обмен часто ограничен рамками отдельного бизнес-процесса (например, подача налоговой декларации) и не предполагает согласованности с другими бизнес-процессами. Обмен осуществляется, как правило, на регулярной основе и позволяет не просто использовать электронные носители информации как альтернативу бумажным, но и производить автоматизированную обработку данных, включая проверку корректности, поиск, учетно-аналитические операции и др. Моделью данных при обособленном взаимодействии можно считать описание структуры электронного документа.

Внутриотраслевое взаимодействие предполагает не просто структурирование данных, но и согласование структур, имен, ограничений и правил контроля корректности для семантически эквивалентных элементов данных в рамках нескольких связанных бизнес-процессов. Уровень сложности структур данных – «унифицированный». Потребность в унификации возникает в ряде сфер деятельности, таких, как таможенное администрирование, транспортно-логистические операции, техническое регулирование и др., где количество используемых форм документов может исчисляться десятками, если не сотнями. Нередко репрезентативный состав этих документов пересекается, содержащиеся в них данные используются на разных этапах бизнес-процессов или в нескольких самостоятельных бизнес-процессах разными участниками. Возникает проблема автоматизированной обработки данных при их дублировании в различных документах: разрешение смысловых

противоречий, различия в размерностях, способах представления и др.

Очевидно решение указанной проблемы состоит в использовании единых согласованных имен, структур, ограничений, правил контроля и других характеристик элементов данных по меньшей мере в рамках отдельно взятой отрасли или предметной области. Здесь как раз и проявляется роль моделей данных как способа представления согласованных характеристик элементов данных. Следует отметить, что в отличие от форм документов модели данных описывают не частные наборы сведений об объектах реального мира, представленные в документах, а объекты реального мира как таковые, представленные набором элементов данных независимо от контекста конкретного документа [7].

Межотраслевое взаимодействие означает переход на принципиально новый характер взаимодействия, когда возникает необходимость в межпредметном, междисциплинарном, межгосударственном обмене данными. Здесь практически неизбежны проблемы совместимости различных моделей данных, построенных в разное время, на разных принципах и правилах, имеющих разный уровень детализации. Зачастую дополнительное отягощение могут придавать факторы, на первый взгляд далекие от ИТ-технологий, такие, как политика, национальный менталитет и др. На момент написания данной статьи общего решения этих проблем не выработано. Конечно общим решением может быть создание единой глобальной модели данных, но в обозримом будущем такое вряд ли возможно. Более реальным выглядит установление соответствия между моделями данных, проведение процедур гармонизации, согласование правил преобразования данных. Здесь весьма перспективным выглядит применение онтологического подхода [8, 9], позволяющего связать разнородные модели данных на уровне онтологических категорий.

Модели данных часто приобретают статус государственных или международных стандартов, например, стандарт электронного обмена в финансовой области ISO 20022 Financial Services – universal financial industry message scheme. Все эти модели, как правило, основаны на разных подходах, принципах и методиках моделирования.

К настоящему времени в различных сферах деятельности разработаны уже десятки моделей данных международного уровня, используемых при информационном взаимодействии. С учетом национальных и ведомственных моделей их количество будет исчисляться сотнями.

О серьезности проблемы совместимости моделей

свидетельствует тот факт, что был разработан и принят комплект стандартов ISO/IEC 19763 Information technology – Metamodel framework for interoperability (MFI), призванный создать основу для обеспечения интероперабельности (совместимости) информационных систем, построенных на различных моделях. Стандарт предусматривает ведение реестра (регистра) моделей (не только моделей данных), а также описывает способ описания отображения (mapping) одной модели на другую.

III. ПАСПОРТИЗАЦИЯ ИНФОРМАЦИОННЫХ МОДЕЛЕЙ

В ходе практической деятельности в поле зрения авторов попало около 20 самых различных информационных моделей, прежде всего моделей данных. Многие из них были изучены достаточно подробно, некоторые поверхностно. Так или иначе, для авторов стала очевидной необходимость в систематизации накопленных знаний.

Тема систематизации и классификации моделей данных отнюдь не нова. К ней неоднократно подходили с различных точек зрения, что находит отражения в многочисленных ИТ-изданиях (в т. ч. сетевых) научного, учебного и научно-популярного характера. Например, в контексте баз данных модели данных обычно классифицируются по характеру отношений между объектами в соответствии с характером СУБД (иерархические, реляционные, сетевые и др.) [10, 11, 12]. Вообще говоря, такую классификацию с некоторыми оговорками можно применять и вне контекста баз данных.

В контексте информационного взаимодействия в рамках социально-экономических процессов возникает естественная классификация моделей данных по широте охвата (межгосударственные, межгосударственные, национальные, корпоративные и др.). Например, в [13], при рассмотрении моделей данных как инструмента импортозамещения в стратегических информационных системах выделяются многоотраслевые, одноотраслевые и системные модели данных.

Тема классификации моделей данных заслуживает самостоятельного исследования и выходит за рамки данной статьи.

При анализе разнородных информационных моделей (не только моделей данных) весьма удобным инструментом является т.н. паспорт информационной модели, построенный на основе набора метаданных Дублинского ядра (Dublin Core, DC), дополненного рядом специфических полей. Перечень полей, описывающих информационные модели, приведен в таблице 1.

Таблица 1. Перечень полей, описывающих информационные модели

Наименование поля	Описание	Поле DC
Название модели	Название модели, используемое ответственной организацией	Title
Ответственная организация	Наименование организации, ответственной за разработку и сопровождение модели	Creator
Предметная область	Предметная область, в рамках которой разработана модель (используются классификаторы ГРНТИ, ОЭСР и др.)	Subject

Описание	Приводится краткое аннотированное описание модели	Description
Дата	Дата выпуска первой официальной версии модели	Date
Тип модели	Тип модели в соответствии с контролируемым словарем (модель данных, модель процессов, архитектурная модель и др.)	Type

Окончание таблицы 1

Наименование поля	Описание	Поле DC
Формат представления	Формат представления модели (PDF, Excel и др.)	Format
Источник	Указываются адрес в сети Интернет (URL) и (или) библиографическое описание (БГО)	Source
Язык	Язык определения элементов модели по стандарту ISO 639-2	Language
Связь с другими моделями	Указываются модели, с которыми связана описываемая модель и характер связей	Relation
Охват	Охват модели с точки зрения уровня масштаба признания в соответствии с контролируемым словарем (международная, межгосударственная, национальная и др.)	Coverage
Права	Юридические аспекты использования модели (особенности распространения, вопросы интеллектуальной собственности и др.)	Rights
Номер актуальной версии	Номер актуальной версии модели	–
Дата актуальной версии	Дата выпуска актуальной версии модели	–
Нотация	Способ описания элементов модели (таблица, UML, BPMN и др.)	–
Стандарт	Обозначение стандартизирующего документа (при наличии)	–
Мета модель	Обозначение методики или метамодели, в соответствии с которой строится модель (при наличии)	–

В рамках заявленного цикла статей будут рассмотрены модели данных, по мнению авторов, наиболее интересные с точки зрения разнообразия использования методических и технологических подходов к разработке, и в то же время наиболее известные. Паспорта этих моделей сведены таблицу 2.

Таблица 2. Учетные данные некоторых моделей данных

Наименование поля	UN/CCL	WCO DM	ISO 20022	NIEM	WIPO
Название модели	The UN/CEFACT Core Component Library (CCL) Библиотека ключевых компонентов CEФАКТ ООН	WCO Data Model Модель данных ВТамО	Репозиторий ISO 20022	U.S. National Information Exchange Model (NIEM) Национальная модель обмена информацией США	Standard WIPO ST.96. Recommendation for the Processing of Industrial Property Information using XML Стандарт ВОИС ST.96. Рекомендации по обработке информации по промышленной собственности с использованием XML
Ответственная организация	UN Centre for Trade Facilitation and E-business (UN/CEFACT) Центр по упрощению процедур торговли и электронным деловым операциям (CEФАКТ ООН)	World Customs Organization (WCO) Всемирная таможенная организация (ВТамО)	ISO/TC 68 Банковское дело и соответствующие финансовые операции	The U.S. Department of Homeland Security Министерство внутренней безопасности США The U.S. Department of Justice Министерство юстиции США	The XML4IP Task Force of the Committee of WIPO Standards (CWS) Целевая группа XML4IP Комитета по стандартам ВОИС
Предметная область	ГРНТИ: 72.23.19 Условия внешне-торговых перевозок; 72.75.49: Стимулирование внешней торговли OECD: 05.02.DI BUSINESS	ГРНТИ: 72.23.19 Условия внешнеторговых перевозок; 72.75.49 Стимулирование внешней торговли OECD: 05.02.DI BUSINESS	Финансовые услуги	Биометрия. Химические, биологические, радиационные, ядерные угрозы. Детство, юность, семья. Чрезвычайные ситуации. Сфера услуг. Иммиграция. Защита инфраструктуры. Разведка. Международная торговля. Правосудие. Морской транспорт. Военные операции. Скрининг. Наземные перевозки	ГРНТИ: 10.35 Патентное право. Право промышленной собственности; 10.41 Авторское право и смежные права; 10.89.35 Международное частное право. Патентное право; 10.89.41 Международное частное право. Авторское право OECD: 05.02.DI BUSINESS; 05.05.OM LAW
Описание	Используется для формирования сообщений в электронной форме при обмене данными между торговыми партнерами в рамках триады «покупка – доставка – оплата».	Используется для формирования сообщений в электронной форме при обмене данными между участниками процессов трансграничной торговли.	Используется для электронного обмена данными между финансовыми институтами.	Используется для межведомственного информационного обмена в США	Используется для подачи, публикации, обработки и обмена информацией по промышленной собственности
Дата	2006	2002	2005	2005	2012
Тип модели	Модель данных	Модель данных Модель процессов	Модель данных Модель процессов Модель сообщений	Модель данных Модель сообщений	Модель данных
Формат представления	Таблицы Excel W3C XML-схемы	Таблицы Excel Текст Диаграммы UML W3C XML-схемы	Escore-модель W3C XML-схемы сообщений	Таблицы Excel W3C XML-схемы Access CSV; XMI	Таблицы Excel W3C XML-схемы
Источник	http://www.unecce.org/cefact/codesfortrade/unccl/ccl_index.html	http://www.wcoomd.org/en/topics/facilitation/instrument-and-tools/tools/data-model.aspx	https://www.iso20022.org/	https://www.niem.gov/	http://www.wipo.int/standards/en/st96/
Язык	eng (английский)	eng (английский)	eng (английский)	eng (английский)	eng (английский)
Связь с другими моделями	Аналог WCO Data Model	Аналог UN/CEFACT CCL	Заменяет: ISO 7775, ISO 15022 Аналог: SWIFT MT; ISO 8583; FIX (Financial Information Exchange); FpML (Financial products Markup Language); XBRL (eXtensible Business Reporting Language)	Основана на GJXDM (Global Justice XML Data Model)	Заменяет: ST.36. Processing patent information using XML; ST.66. Processing trademark information using XML; ST.86. Processing industrial design information using XML.
Охват	Международная модель	Международная модель	Отраслевая модель	Национальная модель	Международная модель
Права	Распространяется	Распространяется на	Распространяется без	Распространяется без	Распространяется без ограничений

	без ограничений	возмездной основе	ограничений	ограничений	
Номер актуальной версии	16B	3.6	Версионность ведется только для сообщений	3.2	2.2
Дата актуальной версии	2016	2016	25.10.2016	15.06.2016	17.08.2016
Нотация	UML	UML	UML	UML	UML
Стандарт	–	–	ISO 20022:2013. Financial services – Universal financial industry message scheme (Parts 1-8)	OMG. UML Profile for NIEM	Standard WIPO ST.96. Recommendation for the Processing of Industrial Property Information using XML
Метамодель	ISO 11179 + CCTS (Техническая спецификация ключевых компонентов)	ISO 11179 + CCTS (Техническая спецификация ключевых компонентов)	Модель Захмана	Архитектура, управляемая моделями	–

IV. ЗАКЛЮЧЕНИЕ

В следующей статье цикла будет подробно рассмотрена Библиотека ключевых компонентов СЕФАКТ ООН, которая, по мнению авторов, является наиболее сложной по организации и, вместе с тем, наиболее технически совершенной среди представленных.

БИБЛИОГРАФИЯ

- [1] Codd E.F. Data Models in Databases Management // Proc. Workshop on Data Abstraction, Database and Conceptual Modelling. – Pingree Park, Colo, June 1980.
- [2] Дейт К. Дж. Введение в системы баз данных, 8-е издание: Пер. с англ. М.: Издательский дом «Вильямс», 2005.
- [3] Дейт К. Дж., Дарвен Хью. Основы будущих систем баз данных. Третий манифест. М.: Янус-К, 2004.
- [4] ГОСТ Р ИСО/МЭК 11179-1–2010. Информационная технология. Регистры метаданных (РМД). Часть 1. Основные положения. М.: Стандартинформ, 2012.
- [5] Tasker D. Fourth Generation Data: A Guide to Data Analysis for New and Old Systems. Sydney, Australia: Prentice-Hall of Australia Pty., Ltd., 1989.
- [6] Арманд А. Д. Иерархия информационных структур мира // Иерархия информационных структур мира // Вестник РАН, 2001, том 71, № 9. С. 797–805.
- [7] Partridge, C.: Business Objects: Re-Engineering for Re-Use. Butterworth Heinemann, 1996.
- [8] Nikiforov, D.A., Lisikh, I.G., Sivakov, R.L.: An approach to multi-domain data model development based on the model-driven architecture and ontologies // Supplementary Proceedings of the 4th International Conference on Analysis of Images, Social Networks and Texts (AIST'2015). Yekaterinburg, Russia, April 9–11, 2015. CEUR Workshop Proceedings, vol. 1452, pp. 106–117. CEUR-WS.org (2015).
- [9] Nikiforov, D.A., Korchagin, A.B., Sivakov, R.L.: An Ontology-Driven Approach to Electronic Document Structure Design // Communications in Computer and Information Science, 5th International Conference on Analysis of Images, Social Networks, and Texts (AIST 2016). Yekaterinburg, Russia, April 7–9, 2016. Vol. CCIS 661. P. 12–25.
- [10] Когаловский М.Р. Энциклопедия технологий баз данных. М.: Финансы и статистика, 2002.
- [11] Кузнецов С. Базы данных. Вводный курс // CIT Forum [Электронный ресурс]. ЦИТ Форум, 2008. URL: http://citforum.ru/database/advanced_intro (дата обращения: 09.02.2017).
- [12] Васильев А.Е. Развитие логических моделей данных // CIT Forum [Электронный ресурс]. ЦИТ Форум, 2010. URL: http://citforum.ru/database/articles/ref_vs_nav_models (дата обращения: 09.02.2017).
- [13] Акаткин Ю., Дрожжинов В., Конявский В. Стандарты моделей данных для обмена информацией как инструмент импортозамещения в стратегических информационных системах // PC Week/RE [Электронный ресурс]. СК ПРЕСС, 2014. URL: <https://www.pcweek.ru/etc/167200.pdf> (дата обращения: 09.02.2017).

Data models for information exchange

A.B. Korchagin, I.G. Lisikh, D.A. Nikiforov, R.L. Sivakov

Abstract – The publication opens a series of articles devoted to the analysis of various approaches to data modeling and revealing their key technological features. This article describes the dependence of the complexity level of data structures from the level of organization of information exchange. The authors also propose the structure of the data model passport and bring examples of passports of the following models: UN/CEFACT CCL, WCO Data Model, ISO 20022, NIEM, WIPO St.96.

Keywords – information exchange; data model; process model; data element; data metamodel; document; message; data model compatibility; data model mapping; ontology.