

Концептуальное моделирование в технологиях баз данных и онтологические модели

М.Р. Когаловский
Институт проблем рынка РАН

Содержание

- Ранние системы баз данных и методы их проектирования
- Трехсхемная технология ANSI/X3/SPARC
- Концептуальное моделирование предметной области
- Инфологический подход к проектированию баз данных
- Роль концептуальной схемы в системах баз данных
- Средства концептуального моделирования
- Концептуальные схемы баз данных и онтологии, сходства и различия
- Терминологические замечания
- Актуальные проблемы концептуального моделирования
- Заключение

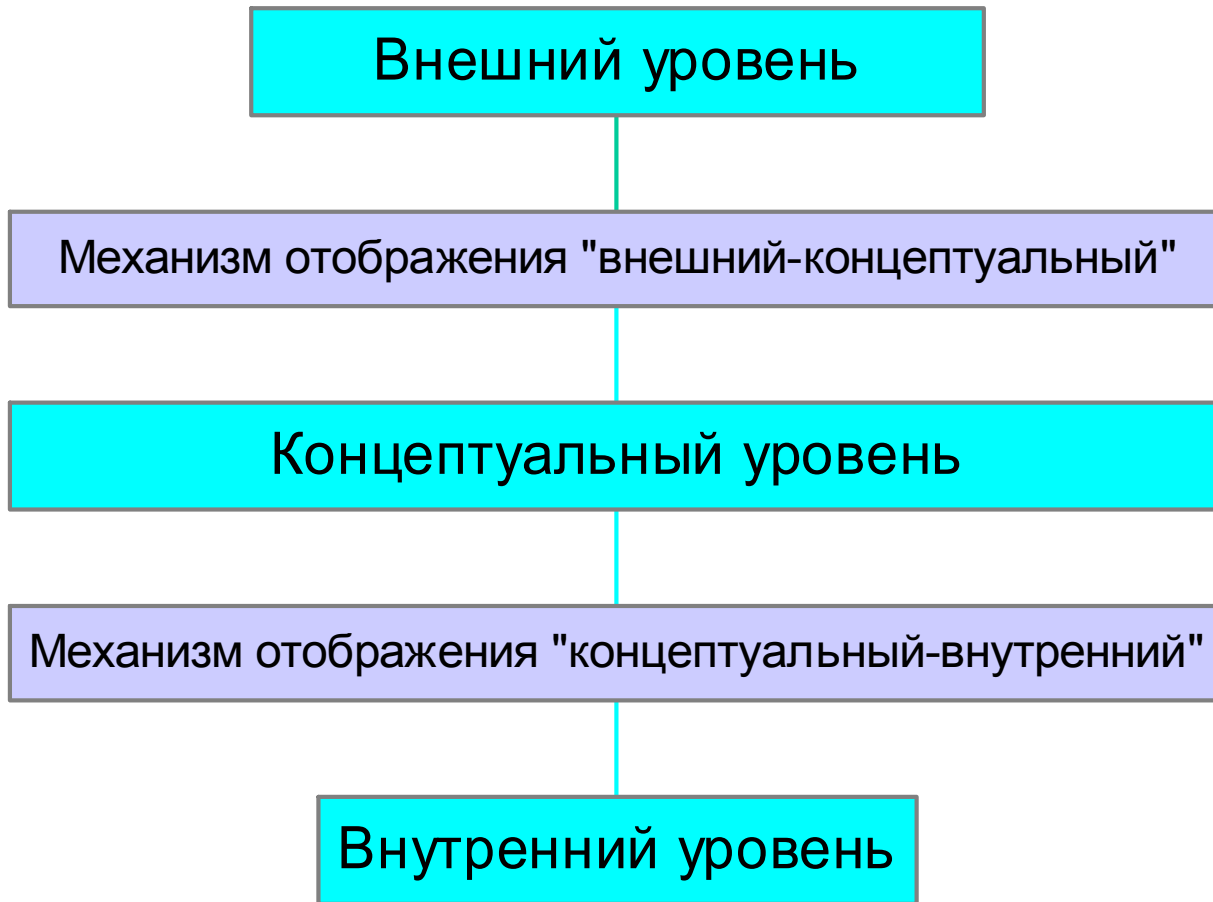
Ранние системы баз данных

- Истоки технологий баз данных
- Ранние БД – технологический элемент приложений
- Двухуровневая архитектура систем БД, независимость данных
- Методы и критерии проектирования ранних БД, ориентированные на реализацию
- Диаграммы Бахмана – прообраз современных средств визуального проектирования
- Рождение индустриального производства СУБД
- Востребованность стандартизации в технологиях БД
- Создание Рабочей группы ANSI/X3/SPARC (1972) (SPARC = The Standards Planning and Requirements Committee)
- Архитектурные исследования – CODASYL, Senko, Nijssen, Palmer и др.

Трехсхемная технология ANSI/X3/SPARC

- Промежуточный отчет Рабочей группы (1975)
- Предложенный подход к стандартизации
- Обобщенная архитектурная модель ANSI/SPARC и независимость данных
- Уровни архитектуры, механизмы междууровневого отображения, обеспечение независимости данных
- Концептуальный уровень, концептуальная схема
- Роль концептуальной схемы в контексте отчета:
 - + интегрированное представление информационных потребностей сообщества пользователей, независимое от реализации
 - + описание контента базы данных
 - + обеспечение разделения данных, содержащихся в базе данных
 - + основа устойчивости архитектуры, управления доступом
- Новый взгляд на этапы проектирования базы данных, концептуальное проектирование.

Архитектура ANSI/X3/SPARC



Что такое концептуальное моделирование?

- Концептуальное моделирование – общая проблема для технологий баз данных, ИИ и программной инженерии
- Наведение мостов:
 - + Joint SIGMOD/SIGPLAN Conf. on Data: Abstraction, Definition and Structure (1976)
 - + The Workshop on Data Abstraction, Databases and Conceptual Modelling (1980)
 - + Монография «On Conceptual Modelling: Perspectives from Artificial Intelligence, Databases and Programming Languages» (1984)
- Концептуальное моделирование (КМ) предметной области (ПО) – моделирование понятий (концептов) и концептуальных структур ПО
- В процессе КМ выявляются, анализируются и описываются релевантные его целям понятия ПО, связи между ними, ограничения, которым они должны удовлетворять
- Цель КМ – синтез концептуальной модели ПО
- Концептуальная модель ПО - это абстрактная модель, независимая от аспектов ее реализации и определяющая концептуальную структуру ПО, релевантную целям моделирования.

Семантика концептуальной модели

- КМ может в различной степени отражать семантику предметной области (соответствовать разным семантическим уровням)
- Семантический уровень концептуальной модели зависит от возможностей выразительных средств, используемых для ее представления
- Лучшими выразительными средствами для этих целей являются естественные языки, но такие модели не структурированы, не обеспечивают однозначной интерпретации
- КМ предметной области, используемые в традиционных технологиях баз данных и в онтологическом моделировании различаются, главным образом, их семантическим уровнем
- Возможности средств представления семантики предметной области в значительной мере связаны с тем, что представляет собой атомарный факт, какие понятия принимаются как предопределенные, какие более крупные информационные структуры образуют осмысленные конструкции
- Разнообразии инструментов концептуального моделирования связано с разными вариантами ответа на эти вопросы.

Концептуальное моделирование в БД (1)

- Концептуальная модель ПО в технологиях БД – это концептуальная схема (КС) базы данных
- Механизмы синтеза КС – абстракции (J.Smith & D.Smith) и иерархии абстракций
- Абстракция состояний и абстракция трансформаций (поведения)
- В области БД главное - абстракции состояний.
Но: объектный подход
- Два вида абстракций: агрегация и обобщение
- Обобщение образует новый объект из класса других объектов
- Агрегация образует объект как отношение между другими объектами.

Концептуальное моделирование в БД (2)

- Назначение КС в технологиях баз данных:
 - + абстрактное определение контента БД достаточно выразительным образом для пользователей, разработчиков и системного персонала
 - + возможность проектирования БД на уровне независимом от специфики используемой СУБД
- Концептуальная схема базы данных – типичная модель предметной области (абстракция обобщения)
- Инструмент представления концептуальной схемы - концептуальная модель данных
- Концептуальной моделью данных называют модель, используемую на стадии концептуального проектирования базы данных
- Концептуальную схему многие отождествляют с концептуальной моделью данных
- Иначе говоря, результат моделирования отождествляют с инструментом моделирования.

Инфологический подход (1)

- Инфологический подход (И.П.) – вклад шведской школы ИС в область КМ , истоки - работы Б. Лангефорша начала 1960-х гг.
- В течение полутора десятилетий серия работ Б. Лангефорша, А. Солвберга, Я. Бабенко и др.
- Адаптация И.П. к технологиям баз данных – Б. Сундгрэн
- И.П. в большей мере, чем традиционные технологии БД, акцентирует концептуальные и семантические аспекты данных
- Предполагается использование моделей данных, которые естественным для человеческого восприятия образом представляют реальность
- Наиболее подходящим для этого выразительным средством являются ЕЯ, однако свойственная им неоднозначность ограничивает их использование для этих целей.

Инфологический подход (2)

- Авторы И.П. предложили в качестве средства инфологического моделирования модель данных, основанную на логике предикатов, расширенной поддержкой концепции времени
- По сути, эта модель представляет собой ограниченную версию естественного языка
- Деятельность по моделированию предметной области при разработке информационных систем и систем баз данных в И.П. разделяется на две сферы – инфологическую и даталогическую
- Разработка ориентированной на человека модели предметной области, независимой от технологий реализации, относится к *инфологической сфере*
- Модель реальности такого рода называется инфологической моделью предметной области, а используемый инструмент моделирования называется *инфологической моделью данных*.

Инфологический подход (3)

- Для целей реализации инфологическая модель предметной области отображается в среду СУБД. Результат отображения называется *даталогической моделью* предметной области
- Таким образом, даталогическая модель предметной области описывает представление инфологической модели ПО в компьютерной среде.
- Сфера моделирования, связанная с созданием и использованием даталогической модели предметной области, называется *даталогической сферой*
- В системах баз данных даталогическая модель предметной области представляет собой *схему базы данных*. С точки зрения архитектурны ANSI/X3/SPARC, это *концептуальная схема базы данных*
- Модели данных, используемые для представления даталогической модели предметной области, естественно называть *даталогическими моделями данных*.

Роль концептуальной схемы в БД (1)

- Публикация отчета ANSI/X3/SPARC вызвала продолжительную дискуссию, какой должна быть КС
- Предмет дискуссий: должна ли КС одновременно выполнять две функции:
 - + функцию модели предметной области, синтезированной в результате ее системного анализа
 - + посредника между пользовательскими представлениями БД и хранимой БД, описывающего контент базы данных
- Другой вопрос: КС для экспертов или для пользователей?
- Отчет ANSI/SPARC: она должна выполнять обе функции, должна иметь машино-читаемую форму и должна быть представлена на хорошо определенном и потенциально стандартизуемом языке
- Т. Стил (рук. Рабочей группы): КС предназначена только для экспертов, и концептуальная модель данных, средствами которой она будет специфицироваться, должна быть дескриптивной моделью, основанной на формальной логике
- Для ее представления должна быть создана концептуальная модель данных, позволяющая удерживать семантику предметной области в КС.

Роль концептуальной схемы в БД (2)

Отчет ISO TC97/SC5/WG3

- Для формирования точки зрения на роль КС в системах БД , определения концепций и терминологии для языков КС с целью обеспечения основы для стандартизации в данной области в 1977г. была создана Рабочая группа ISO TC97/SC5/WG3
- В 1982 г. Рабочая группа выпустила отчет:
Concepts and terminology for the conceptual schema and the information base. Ed. by J.J. van Griethauzen. ISO TC97/SC5/WG3.
- Аналогично подходу ANSI/SPARC, на КС в отчете возлагаются две функции:
 - (1) это - средство обеспечения устойчивого базиса для отображения между внешними и внутренней схемами базы данных
 - (2) это - основа коммуникации между причастными к системе базы данных людьми.
- В отчете предложена терминология КС
- Для иллюстрации требований к выразительным средствам КС рассмотрены: ER-модель; модель, основанная на логике предикатов; модели бинарных и n-арных отношений; подробно рассмотрена модель данных «объект-роль» Фалкенберга.

Роль концептуальной схемы в БД (3)

Отчет ISO TC97/SC5/WG3

- Для выполнения указанных функций нужно иметь языки описания КС в форме, «понятной» компьютеру, и в форме, понятной человеку
- Эти роли может выполнять как один, так и разные языки
- Механизмы поддержки КС должны обеспечивать легкую ее модификацию с тем, чтобы отражать те изменения, которые происходят в предметной области
- Представления информации о предметной области для различных пользователей не должны противоречить декларациям КС (согласованность внешних схем и КС)
- Наконец, КС должна быть устойчивым инвариантом относительно возможных изменений во внутренней схеме базы данных
- Эти требования к КС рассматриваются как минимум, которому должна удовлетворять любая модель данных, претендующая на роль концептуальной модели данных
- Выпуском отчета TC97/SC5/WG3 деятельность ISO по стандартизации концептуальной модели завершилась.

Роль концептуальной схемы в БД (4)

- Ряд специалистов (У. Кент и др.) высказался за два языка КС :
 - + Расщепить КС на два слоя – на описание ПО и коллективную (центральную) схему
 - + Описание ПО - ее семантическая концептуальная модель
 - + Коллективная схема – описание контента БД
 - + Семантическая модель ПО относится к инфологической сфере , используется лишь на стадии проектирования БД
 - + Коллективная схема относится к даталогической сфере, используется и на стадии проектирования, и на стадии функционирования системы БД
 - + Итак, должны существовать две взаимосвязанные версии концептуальной модели ПО:
 - (1) КМ, которую можно назвать КС предметной области
 - (2) КС базы данных в смысле отчета ANSI/SPARC, порождаемая отображением первой в среду используемой СУБД.

Средства концептуального моделирования

- Языки описания концептуальных моделей предметной области в технологиях баз данных воплощают функциональность дескриптивных средств концептуальных моделей данных
- Эти модели данных не зависят от реализации проектируемой системы
- Примеры таких языков - ER-диаграммы, диаграммы классов UML + язык OCL, язык ODL стандарта ODMG и др.

Концептуальные модели данных (1)

- В процессе эволюции технологий баз данных был разработан целый ряд концептуальных моделей данных:
 - + модель «сущностей-связей» (ER-модель) Чена
 - + модель «объектов-ролей» ORM Фалкенберга
 - + ее прототип – методология моделирования NIAM Найсена (Nijssen's Information Analysis Methodology)
 - + модели, основанные на бинарных семантических сетях - модель данных Абриаля, которая интегрирует идеи семантических сетей и языков программирования,
 - + язык FORAL Сенко
 - + модель Браччи, Паолини и Пелагетти.

Концептуальные модели данных (2)

- Другие модели семантического уровня:
 - + расширенная реляционная модель RM/T Кодда,
 - + функциональная модель DAPLEX Шипмана
 - + семантическая реляционная модель Шмида и Свенсона,
 - + модель SDM Хаммера и Маклеода [SDM]
 - + инфологическая модель Бо Сундгрена
 - + модель Taxis Милопулоса и др.
- Были разработаны также подходы:
 - + IAM Бабенко
 - + подход Бахмана и Дайя к концепции роли в моделях данных и многие другие
- Язык SYNTHESIS
- Высокоуровневые модели данных называют семантическими моделями
- Работы D. Smith & D. Smith по методам абстракции данных.

Проект TAXIS (1)

- Язык TAXIS – воплощает концептуальную модель данных высокого семантического уровня для концептуального моделирования предметной области
- Исследовательский проект на стыке ИИ и технологий баз данных, начало 80-х гг., университет Торонто, Дж. Милопулос и др.
- Основан на объектном подходе и семантических сетях
- Объекты могут обладать свойствами/атрибутами
- Объекты организуются в классы, классы сами являются объектами 1-го класса и могут быть членами метаклассов
- Поэтому классы могут также обладать собственными свойствами
- Поддерживается иерархия классов и отношения специализации классов и наследования свойств объектов.

Проект TAXIS (2)

- В TAXIS имеются возможности для моделирования активностей: транзакции (кратковременные активности) и скрипты (долговременные активности)
- Скрипты строятся на основе сетей Петри, узлами которых являются транзакции
- Предусмотрены средства описания пользовательских интерфейсов на основе объектной парадигмы
- Имеется возможность моделирования требований к создаваемой системе: язык моделирования требований Requirement Modeling Language (RML)
- Разработана методология и инструментальная среда проектирования баз данных и информационных систем.

Отечественные проекты

- Влияние инфологического подхода на отечественные разработки в области проектирования баз данных
- В 1980-е гг. в СССР был выполнен ряд исследовательских проектов, связанных с созданием высокоуровневых концептуальных моделей данных , например Михновский (1983), Вейнеров и др.(1985), Фурсин (1987), Замулин (1990)
- В ИПИ РАН разработан язык высокого уровня СИНТЕЗ (Калиниченко, 1993)
- СИНТЕЗ воплощает семантическую модель данных для концептуального моделирования предметной области, позволяющий специфицировать и концептуальную схему предметной области и активности
- Язык развивается и поддерживается до настоящего времени и используется в качестве базового языка в ряде исследовательских проектов.

Индустриальные технологии (1)

- Сложившиеся индустриальные технологии КМ основаны, главным образом, на двухверсионности КС
- В индустриальных технологиях КМ доминируют методы ОА&D и объектные модели данных, UML и ER-модель
- Язык UML с конца 1990-х гг. при поддержке OMG быстро стал стандартом де-факто
- UML поддерживается большинством коммерческих CASE-инструментов
- В CASE-инструментах, поддерживающих язык UML, концептуальная модель предметной области специфицируется в терминах диаграммы классов UML с комментариями или OCL-декларациями
- Язык OCL (Object Constrain Language) стандарта UML служит для декларации ограничений, налагаемых на элементы диаграммы классов.

Индустриальные технологии (2)

- Построенные в среде CASE-инструментов спецификации отображаются в даталогическую среду, генерируя КС базы данных на языке SQL, ODL стандарта ODMG или др. языках описания данных концептуальной модели данных, которая поддерживается используемой СУБД
- Сгенерированные «машино-ориентированные» КС баз данных поддерживаются механизмами используемых СУБД
- Поддержка инфологических КС механизмами СУБД не обеспечивается индустриальными технологиями
- В последние годы оказывают влияние разработки SW
- Научные системы - RDF, RDFS и OWL, программные средства инженерии онтологий
- Специальные разработки, например, расширение языка XML – Conceptual XML (C-XML).

Концептуальные схемы и онтологии: сходства

- Оба эти артефакта - порождаются в результате концептуального моделирования предметной области
- Как онтология, так и концептуальная схема, являются частичными описаниями концептуализации предметной области, представляющими структуру и ограничения (правила) моделируемой предметной области
- И онтология, и концептуальная схема являются *абстрактными* моделями предметной области, представляющими различные по характеру представления ее концептуальной структуры в форме, независимой от реализации систем, в которых они используются
- И онтологии, и концептуальные схемы используются для семантического моделирования предметной области в различных системах, в системах семантической интеграции информации из множества источников и обеспечения коммуникаций в предметной области
- Одни и те же языки концептуального моделирования часто используются для представления онтологий и КС.

Концептуальные схемы и онтологии: различия - 1

- В обоих случаях моделируется «предметная область», но в англоязычной литературе по БД используют термины Enterprise или Application Domain, а в литературе по онтологическому моделированию – Universe of Discourse или Subject Domain
- Термины БД подчеркивают ограниченность ПО информационными потребностями конкретных приложений или конкретного сообщества пользователей
- В отличие от КС при разработке онтологии стремятся как можно к большей степени ее общности и независимости от информационных потребностей отдельных приложений или групп пользователей. Поэтому онтология включает относительно общие знания
- Онтология может повторно использоваться различными приложениями, КС не предназначена для повторного использования
- В КС традиционных систем БД и в онтологиях используются разного рода семантические отношения между понятиями. В КС ограничиваются, как правило, отношениями агрегации и обобщения. В онтологиях более богатый набор конкретизированных семантических отношений
- Даталогические КС зависят от реализации систем БД, а онтологии являются абстрактными моделями предметной области, независимыми от реализации использующих их систем.

Концептуальные схемы и онтологии: различия - 2

- КС в технологиях БД оперирует только типами сущностей (классами) и типами связей, а некоторые онтологии имеют дело и с индивидами
- КС, как и онтология, описывает концептуальную структуру ПО, но онтология должна включать еще определения используемых понятий
- Онтология – это теория ПО, а КС базы данных – описание ее концептуальной структуры
- Семантика КС - результат неформального соглашения между ее разработчиками, пользователями и системным персоналом. Онтология – теория ПО, являющаяся результатом консенсуса достаточно широкого сообщества относительно единой трактовки смысла описываемых понятий
- КС инфологической сферы используется в традиционных технологиях БД только на стадии проектирования базы данных. Онтология используется на стадии исполнения
- КС в традиционных технологиях БД строится на основе гипотезы замкнутого мира (Close World Assumption), онтология – напротив, на гипотезе открытого мира (Open World Assumption), как в системах, основанных на знаниях
- Сфера применения онтологий шире, чем КС баз данных. Онтологии используются не только в области ИС и систем БД, но и в инженерии знаний, в программной инженерии и в других областях информационных технологий.

Терминологические некорректности

- *Концептуальную схему БД называют онтологией ПО*
- *Схему БД называют моделью данных*
- *Модель данных и модель базы данных рассматриваются как синонимы, и этими терминами обозначают структуру БД (определяемую схемой БД); эта трактовка «перекочевала» в стандарты XML*
- *Модель ПО отождествляется с моделью данных (результат vs инструмент моделирования)*
- *Концептуальная модель ПО отождествляется с концептуальной моделью данных*
- *Концептуальная схема БД отождествляется с концептуальной моделью данных*
- *Эллипсис наименований терминов, например, термин концептуальная модель без указания объекта моделирования. Но существуют концептуальные модели ПО и концептуальные модели данных.*

Актуальные проблемы концептуального моделирования (1)

- Хотя КМ в технологиях БД представляет собой продвинутую область, исследования в этой области продолжаются
- Авторитетным центром концентрации интересов и форумом для обсуждения результатов, полученных в области концептуального моделирования, повидимому, остается стартовавшая в 1979 году международная конференция *Entity-Relationship Approach to Information Modeling and Analysis (ER)*. В 2007 г. состоялась уже 26-я конференция этой серии
- Конференция ER первоначально планировалась как форум для обсуждения направлений развития и практического применения предложенной П. Ченом (1976) в качестве концептуальной модели данных сегодня широко известной модели «сущностей-связей»
- Однако позднее, в 1996 г., название конференции изменилось для приведения его в соответствие с расширившейся ее тематикой. Новое название явно обозначает направленность на проблематику концептуального моделирования
- Главная часть нового названия конференции теперь - **Conceptual Modeling – ER**.

Актуальные проблемы концептуального моделирования (2)

- Актуальные проблемы КМ (по тематике конференций ER):
 - + Взаимосвязь концептуального и онтологического моделирования
 - + Концептуальное моделирование, управляемое онтологией
 - + Разработка онтологий на основе концептуальной схемы
 - + Объектный подход в концептуальном моделировании
 - + Оценка качества и метрики концептуальных моделей, в частности, определение метрик для UML-диаграмм
 - + Связь технологий концептуального моделирования и технологий SW
 - + Методы извлечения информации для формирования концептуальной модели предметной области из текстовых и XML-документов, например, извлечение информации о концептуальных связях
 - + Подходы к КМ для нетрадиционных и актуальных областей применения, например, электронный бизнес, электронное обучение, науки о жизни, разработка веб-приложений, создание систем, оперирующих темпоральными и/или пространственными данными, ГИС и системы пространственных данных, системы с агентной и веб-сервисной архитектурой
 - + Инструментальные проблемы, лучшая практика использования UML, XML-технологии в КМ: расширение для этих целей языка XML, использование спецификаций RDF и OWL.

Заключение

- Концептуальное моделирование предметной области в технологиях баз данных – важный этап процесса проектирования баз данных
- В этой области продолжаются активные исследования и инструментальные разработки, направленные на создание более эффективных технологий разработки систем баз данных и информационных систем
- Разработка концептуальной схемы базы данных в технологиях баз данных и создание онтологии предметной области – близкие виды деятельности.
- Хотя эти артефакты обладают некоторым сходством, между ними имеются также рассмотренные выше существенные различия, что приводит в ряде случаев к подмене понятий.

Литература

1. ANSI/X3/SPARC Study Group on Data Base Management Systems. Interim Report. SIGFDT Bull., ACM, v. 7, no. 2 (1975).
2. Brodie M.L., Mylopoulos J., Schmidt J.W. (eds.): On Conceptual Modelling, Perspectives from Artificial Intelligence, Databases, and Programming Languages. Springer-Verlag, 1984.
3. ISO TC97/SC5/WG3. Concepts and terminology for the conceptual schema and the information base. Ed. by J.J. van Griethauzen. 1982.
4. Peter Pin-Shan Chen. The entity-relationship model - toward a unified view of data. ACM TODS, v.1, n.1, March 1976.
5. Smith J.M. and Smith D.C.P. Database Abstraction: Aggregation and Generalization. ACM TODS, 2(2): 1977.
6. Sundgren B. An infological approach to data bases. – Stockholm: National Central Bureau of Statistics, 1973.
7. Цикритзис Д., Лоховски Ф.. Модели данных /Пер. с англ. – М.: Финансы и статистика, 1985.
и др. (всего 63 источника - см. текст доклада).

Спасибо за внимание.
Вопросы?