2. Основные понятия

2.1. Данные и модели данных

Стремление к приобретению знаний присуще самой природе человека. Наиболее эффективный способ получения новых знаний – проведение научных исследований. Вместе с тем любой индивидуум, наблюдая некоторое явление, получает определенное приращение знаний, которое мы будем называть информацией. Информация может обладать большой ценностью и подлежать регистрации для обеспечения ее доступности множеству заинтересованных лиц. Достижение этой цели существенно связано с представлением информации.

Изначальным средством представления информации служит естественный язык. Однако это не всегда наилучшее средство. Во-первых, естественный язык в силу своей универсальности в некоторых случаях менее эффективен, чем специализированные средства. Во-вторых, естественный язык ориентирован на вольный стиль общения и не обеспечивает необходимой точности регистрации и передачи информации. И, наконец, в силу свойств естественного языка обработка созданных с его помощью текстов на компьютере связана с принципиальными трудностями.

Наряду с решением сложных вычислительных задач компьютеры используются как интеллектуальные помощники, обеспечивающие управление и манипулирование информацией. Вероятно, это останется главной функцией компьютера и в перспективе. Эффективность ее реализации зависит от способов компьютерного представления данных и информации.

Известен ряд различных способов организации данных (таблицы, списки, формы и т.д.). Проблематика моделирования данных связана с таким представлением данных, которое наиболее естественно отражает реальный мир и может поддерживаться компьютерными средствами. Одним из самых мощных и эффективных инструментов, предназначенных для решения указанных задач, в настоящее время являются базы данных.

В базе данных, несомненно, хранятся данные о ПрО.

Определение 2.1.1. Данные (лат. data, от ед.ч. datum – факт) – это факты реального мира и идеи, представленные в формализованном виде, позволяющем передавать или обрабатывать их при помощи некоторого процесса и соответствующих технических средств.

Однако, такие данные, как 120 и 30 ничего не говорят человеку. Его интересует информация.

Информация (от лат. informatio – разъяснение, изложение), первоначально – сведения, передаваемые одними людьми другим людям устным, письменным или каким-либо другим способом. Однако, в связи с «информационным взрывом» во второй половине XX века, произошли изменения в трактовке понятия «информация». Оно было расширено и включило обмен сведениями не только между человеком и человеком, но также между человеком и автоматом, автоматом и автоматом, а кроме этого – обмен сигналами в животном и растительном мире.

В нашей ситуации обмена информацией между человеком и БД можно дать следующее определение.

Определение 2.1.2. Информация – приращение знаний человека, которое может быть получено на основе данных.

Изучением процесса получения информации из данных в знаковых системах занимается наука семиотика.

Определение 2.1.3. Семиотика (греч. semeiotikon, от semeion – знак, признак) – комплекс научных учений, изучающих свойства семиотических (знаковых) систем, основное назначение которых – выражать некоторое содержание.

Термин «знак» понимается в широком смысле как некоторый объект (вообще говоря, произвольной природы), которому при определенных условиях (образующих в совокупности знаковую ситуацию) сопоставлено некоторое значение, могущее быть конкретным физическим предметом (явлением, процессом, ситуацией) или абстрактным понятием.

Процесс получения информации в знаковых системах проходит в три этапа, каждому из которых соответствует свой раздел семиотики.

Определение 2.1.4. Синтактика – раздел семиотики, изучающий внутренние свойства систем знаков безотносительно к интерпретации (синтаксис – правила построения знаков и знакосочетаний в рамках знаковой системы).

Определение 2.1.5. Семантика – раздел семиотики, рассматривающий отношение знаков к обозначаемому (содержание знаков) или, что то же, соотношения между знаками и их интерпретациями, независимо от того, кто служит «адресатом» (интерпретатором).

Термин «семантически значимый», который мы в дальнейшем будем часто использовать, предполагает значимость с точки зрения семантики, изучающей семантические отношения, которые образуются между объектами и знаками, представляющими эти объекты в знаковой системе. Таким образом, семантически значимыми могут быть как объекты (если они представлены некоторыми знаками в семиотической системе), так и знаки (если они определяют некоторые реальные объекты ПрО).

Определение 2.1.6. Прагматика – раздел семиотики, изучающий восприятие выражений знаковой системы в соответствии с разрешающими способностями воспринимающего. Прагматика исследует связь знаков с «адресатом», т. е. проблемы интерпретации знаков теми, кто их использует, их полезность и ценность для интерпретатора.

Поскольку знак есть носитель информации, семиотика получает большое прикладное значение при исследовании и проектировании знаковых систем, используемых в процессах передачи и обработки информации.

Весьма упрощенно процесс получения информации из данных можно проиллюстрировать примером, начатым выше. Нам предъявлены данные – 120 и 30. С точки зрения синтаксиса это целые положительные числа, представленные арабскими цифрами в 10-тичной системе счисления. На этапе синтаксического анализа никакие семантические отношения с объектами ПрО не возникают.

Если эти данные снабдить интерпретацией «*первое число – рост человека в сантиметрах, второе – возраст этого же человека в годах*», можно образовать содержание (смысл) сообщения в целом. Получается, что некий человек имеет значение характеристики *Рост*, равное 120, и значение характеристики *Возраст*, равное 30. На этом завершается этап семантического анализа сообщения и начинается работа прагматики.

Окончательный результат – полученная информация, как приращение знаний человека, являющегося адресатом сообщения, существенным образом зависит от тех знаний, которые он имел на момент получения сообщения.

Человек, не имеющий твердых представлений о типичных значениях указанных характеристик, узнает лишь то, что такие люди тоже встречаются. Опытный человек почувствует что-то нетипичное и будет строить возможные гипотезы: карлик, горбун, человек с ампутированными ногами, … Без дополнительной информации любое из этих предположений может оказаться верным. При наличии сведения, что человек, о котором идет речь в сообщении, живет в Африке, на первый план выйдет гипотеза о том, что он – пигмей.

Последний вывод и является информацией, полученной адресатом сообщения. Приведенный пример намеренно усложняет последний этап анализа сообщения, на котором работает прагматика. Чаще все гораздо проще, и то новое, что узнает

пользователь в диалоге с системой БД, и является информацией. Главное, что должен был продемонстрировать пример, – данные без интерпретаций «мертвы».

В естественном языке и данные, и их интерпретации указываются совместно в выражениях языка. В информационных системах данные всегда долговременно хранятся на диске. А вот в зависимости от того, где находятся их интерпретации, можно выделить три класса систем:

1. Интерпретациями владеет только человек.
2. Интерпретации указаны в программе.
3. Интерпретации также хранятся на диске в виде специальным образом организованных данных.

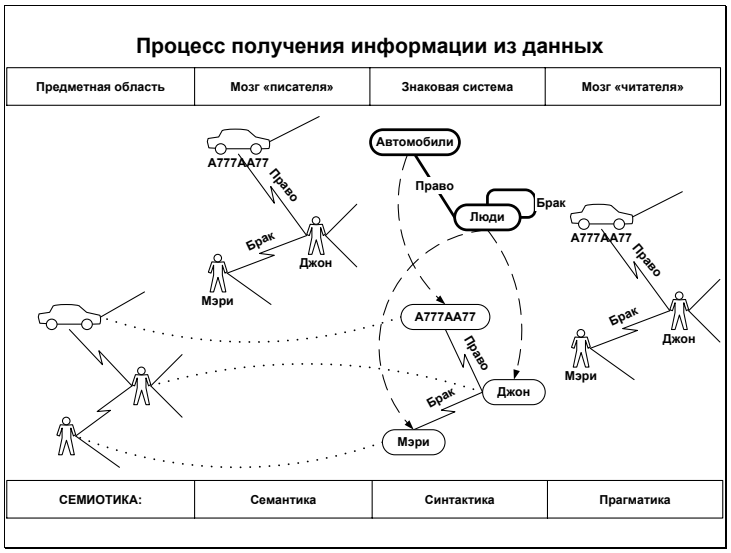
Типичным примером системы первого класса является обычный калькулятор. Действительно, только человек знает, что он складывает – рубли или яблоки – и как интерпретировать полученный при этом результат.

Следующий класс систем можно пояснить на примере программы печати ведомости на выдачу зарплаты. В файле на диске хранится массив строк с фиксированным форматом: первые 20 символов содержат фамилию работника, следующие 5 символов – значение его зарплаты. Если просто распечатать этот файл, получим неинтерпретированные данные. Характерные сочетания первых 20-ти символов, скорее всего сами подскажут человеку, что речь идет о значениях характеристики *Фамилия*. Но о чем говорят вторые элементы, без дополнительных подсказок наверняка не ясно.

Специально написанная программа с помощью операций печати соответствующих констант придает отчету стандартную форму ведомости на зарплату: заголовок, шапка таблицы, завершающие подписи и т.д. Кроме этого, в ней организуется цикл чтения очередной записи из файла и печать ее содержимого в виде строки в теле таблицы. В результате получается отчет, не вызывающий никаких сомнений у кассира.

По такому принципу строили ИС в доСУБДэшную эпоху – управление данными в файле осуществлялось специализированными программами (для каждого файла свои программы). На смену им пришли СУБД – универсальные программные комплексы, предназначенные для управления БД о любых ПрО. Это стало возможным благодаря тому, что в них интерпретации не фиксируются в программах, а хранятся на диске, как и сами данные. Именно их мы задаем на этапе проектирования схемы БД. Заданные однократно интерпретации ассоциируются с данными при их вводе, и в дальнейшем при всех манипуляциях с данными эта связь остается неразрывной, обеспечивая, таким образом, получение информации.

В результате такой эволюции изменилась роль данных. Их уже нельзя рассматривать как совокупность битов, они приобретают определенную интеллектуальную окраску. В таком качестве их можно расценивать как семантически значимое представление реального мира, как взгляд на мир.



Поясним, как осуществляется передача информации о ПрО в технологии БД.

В ПрО независимо от нашего сознания существуют некоторые объекты, обладающие характеристиками и вступающие в связи с другими объектами. На схеме в левой части показаны два объекта типа *ЧЕЛОВЕК*. Это супруги *Джон* и *Мэри*. Изображен также автомобиль, владельцем которого является *Джон*. Соответствующие связи показаны в виде зигзагообразных дуг. Человек, чьей обязанностью является наблюдение за ПрО и отражение всех изменений в БД (назовем его «писателем»), так или иначе, узнал о существовании этих объектов и связей. При этом он произвел у себя в голове первичную формализацию ситуации, в частности, определил знаки для идентификации объектов и связей (на схеме зрительные образы объектов сопровождают эти знаки).

После этого «писатель» обратился к знаковой системе – системе БД. В ее схеме он обнаружил подходящую подсхему (она представлена графом типов в верхней части столбца «Знаковая система») и создал в БД новые объекты типов *ЛЮДИ* и *АВТОМОБИЛИ*, а также связи типов *БРАК* и *ПРАВО*. Получился подграф знаков, изображенный в нижней части того же столбца. Пунктирные направленные дуги связывают знаки с их типами. Именно эти связи указывают, как интерпретировать соответствующие данные. На этом работа «писателя» закончена.

Теперь по мере необходимости любой «читатель» (человек, желающий получить от системы БД информацию о ПрО) может воссоздать представления «писателя» о ПрО, не прибегая к ее непосредственному исследованию. Вместо этого он обращается к системе.

В нижней части схемы показаны разделы семиотики, участвующие в описанных процессах. Синтактика определяет правила построения БД. Семантика изучает семантические отношения между знаками и объектами реального мира (они показаны точечными дугами). И, наконец, прагматика осуществляется в голове у «читателя».

Большие возможности организации и представления информации в компьютеризированных системах обеспечиваются применением описываемых в настоящей книге моделей данных. Любая СУБД в состоянии обеспечить указанные процессы, благодаря тому, что каждая из них поддерживает некоторую свою модель данных. Первое определение этого понятия мы дадим с точки зрения их назначения.

Определение 2.1.7. (Функциональное определение модели данных). Модель данных (МД) – это интеллектуальное средство, позволяющее реализовать интерпретацию данных и таким образом способствующее получению информации. Громкий эпитет «интеллектуальное» применен в данном случае неспроста – способность к восприятию, хранению и передаче информации – одно из важнейших свойств интеллекта.

Атомарная единица информации

<Идентификатор объекта, Наименование признака, Значение признака, [Время]>

<Джон, Рост в см, 180> **<Джек, Рост в см, 190>**

<Джон, Вес в кг, 90> **<Джек, Вес в кг, 80>**

Рост

ID_объекта	Значение
Джон	180
Джек	190

Вес

ID_объекта	Значение
Джон	90
Джек	80

ЧЕЛОВЕК

ID_объекта	Рост	Вес
Джон	180	90
Джек	190	80

Как вы уже поняли, основными объектами исследований в моделировании данных являются данные и их интерпретации. Эта область человеческих знаний, как и многие другие, имеет свой элементарный объект – **атомарную единицу информации (АЕИ)**. Она определяется четверкой – *<Идентификатор объекта, Наименование признака, Значение признака, [Время]>*.

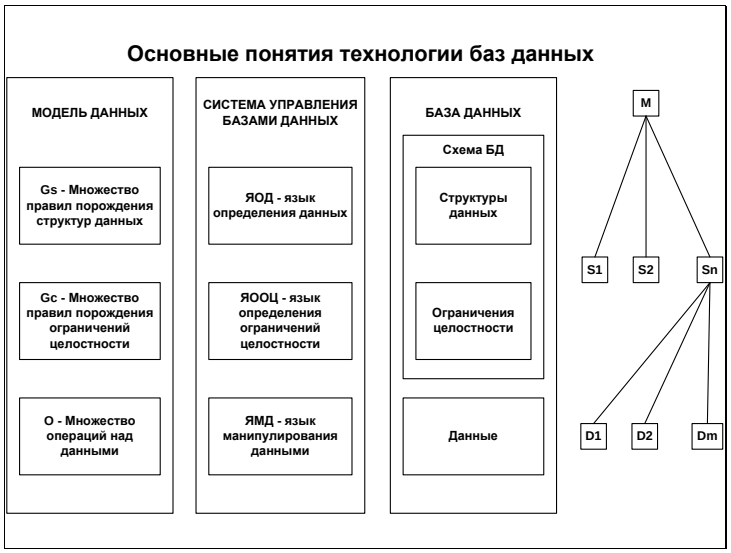
Каждая АЕИ задает истинность следующего факта: «объект, на который указывает идентификатор, имеет определенное значение признака, заданного именем, в конкретное время». Если иметь в виду, что к признакам объектов относятся и свойства, и характеристики, и отношения, понятно, что сколь угодно сложные ПрО можно представить множеством таких четверок.

Последний элемент четверки не напрасно заключен в квадратные скобки, означающие, что он может быть опущен. Действительно, все остальные элементы являются обязательными. Если опустить хотя бы один из них, информация не образуется. Что же касается времени, то не существует моделей данных, в которых механизмы работы с так называемыми темпоральными данными были бы доведены до идеала. Более того, подавляющее большинство моделей и БД предполагают хранение информации об одном (текущем) состоянии ПрО. Такие БД еще называют оперативными, подчеркивая тот факт, что они меняются синхронно с изменением состояния ПрО. Предыдущие состояния данных в них не сохраняются. Поэтому в дальнейшем под АЕИ будем понимать тройку – *<Идентификатор объекта, Наименование признака, Значение признака>*. Примеры АЕИ приведены на слайде.

Разнообразие способов представления атомарных единиц информации, их взаимосвязей и агрегатов породило множество моделей данных.

Проиллюстрируем сказанное на примере синтеза простейшей модели данных. Одним из недостатков атомарной модели данных является многократное дублирование наименований признаков и идентификаторов объектов.

Последовательно избавимся от этой избыточности. Сначала построим таблицы признаков (*Рост* и *Вес*), а затем соберем в одну таблицу значения всех признаков однотипных объектов. Полученная таблица описывает все объекты одной **категории** – *ЧЕЛОВЕК*. Понятие категории является основным структурным понятием одноименной модели – **категориальной модели**. Она предполагает разбиение всех объектов ПрО по категориям. Для каждой категории определяется набор признаков, значения которых характеризуют объекты данной категории. На примере этой простой модели познакомимся с основополагающими понятиями моделирования данных.



Определение 2.1.8. Совокупность именованных категорий и их признаков, а также ограничений на допустимые данные называется **схемой БД**.

Определение 2.1.9. Совокупность данных, структура и значения которых соответствуют конкретной схеме, называется **базой данных (БД)**.

По поводу этих определений следует сделать несколько замечаний.

Во-первых, как мы увидим в дальнейшем, каждая модель данных предполагает свой набор понятий, используя которые мы структурируем ПрО. Наши определения даны для категориальной модели, поэтому в них используются понятия «категория» и «признак». Тем не менее, лучше использовать конкретные, но конструктивные определения, чем давать «размытые» определения типа «база данных – это некоторый набор перманентных данных, используемых прикладными системами какого-либо предприятия».

Во-вторых, в этих определениях указано основное содержание понятий «схема» и «база данных». Иногда это содержание так или иначе расширяют. Так в схему БД могут быть включены программы обработки данных, определения запросов, диалоговых форм и т.д. БД тоже может представляться как все, что хранится на диске помимо собственно СУБД. Мы и сами иногда будем поступать так. В частности, на приведенной схеме БД вобрала в себя помимо собственно данных еще и их схему.

Теперь мы можем взглянуть на модель данных изнутри.

Определение 2.1.10. (Структурное определение модели данных). Модель данных (МД) определяется двумя множествами G и O. G – множество правил порождения схем, O – множество операций над данными. В свою очередь во множестве G выделяются два подмножества – G_S (правила порождения структур данных) и G_C (правила порождения ограничений целостности).

Поясним указанные компоненты моделей данных. Для нашей категориальной модели правила множества G_S могут выглядеть следующим образом:

- БД – это совокупность таблиц.
- Каждая таблица предназначена для хранения информации об объектах одной категории. Имя таблицы – это имя категории.
- Для каждой категории определяется набор признаков, представляющих интерес для объектов этой категории. Имена признаков составляют шапку соответствующей таблицы.
- Каждый объект категории представляется в виде строки таблицы, в столбце признака указывается его значение для данного объекта.

Мы пока детально не обсуждали ограничения целостности, накладывающие дополнительные условия на вводимые в БД данные, тем не менее, приведем несколько правил множества G_c.

Допустимые значения признаков можно ограничить:

- указанием их типа (символьные, числовые, даты и т.д.),
- перечислением этих значений,
- сравнением значений с константой.

Множество операций над данными *O* может выглядеть для нашей модели так:

- операция *INSERT* для добавления новой строки в таблицу,
- операция *UPDATE* для изменения значений одного или нескольких признаков в строке таблицы,
- операция *DELETE* для удаления строки из таблицы,
- операция *SELECT* для поиска строк таблицы, удовлетворяющих определенному условию.

Правила множества *G* обеспечивают порождение схем S_i, каждая из которых определяет конкретную структуру данных и ограничения целостности. Каждой из таких схем в разные периоды времени могут соответствовать различные состояния БД D_j.

Для нашего примера структурный компонент схемы может определять следующие категории или таблицы (здесь вслед за именем категории перечисляются имена признаков объектов этой категории):

- *ЧЕЛОВЕК (Фамилия, Имя, Отчество, Возраст, Пол, Рост, Вес, ...)*
- *АВТОМОБИЛЬ (ГосНомер, Марка, Фирма, ДатаВыпуска, ЦветКузова, ...)*

Используя указанные правила, зададим следующие ограничения целостности:

- Значения признаков *Фамилия, Имя, Отчество, Пол* имеют тип *СТРОКИ*.
- Значения признаков *Возраст, Рост, Вес* имеют тип *ЦЕЛЫЕ ЧИСЛА*.
- Значения признака *ДатаВыпуска* имеют тип *ДАТЫ*.
- Признак *Пол* имеет два допустимых значения '*мужской*', '*женский*'.
- Признак *Возраст* принимает значения, большие 0 и меньшие 150.

Состояние БД может быть следующим.

ЧЕЛОВЕК

Фамилия	Имя	Отчество	Возраст	Пол	Рост	Вес	…
Иванов	Иван	Иванович	50	мужской	180	80	…
Петров	Петр	Петрович	30	мужской	185	85	…
Попова	Мария	Олеговна	40	женский	165	70	…

АВТОМОБИЛЬ

ГосНомер	Марка	Фирма	ДатаВыпуска	ЦветКузова	…
А111АА70	2101	ВАЗ	12.12.1980	белый	…
В222ВВ70	966	ЗАЗ	13.12.1975	черный	…

Определение 2.1.11. Управление БД на ЭВМ осуществляется специализированными программными средствами – **системами управления базами данных (СУБД)**, каждая из которых предлагает свои языковые и диалоговые формы для множеств *G* и *O*: язык определения данных (ЯОД) и язык манипулирования данными (ЯМД). Иногда выделяют отдельный язык определения ограничений целостности (ЯООЦ), но чаще ограничения целостности задаются вместе со структурой в командах ЯОД.

Все приведенные здесь основные понятия технологии БД тесно взаимосвязаны и образуют целостную систему, представленную на последнем слайде. Любое моделирование данных невозможно без использования той или иной модели (левый

столбец). Для этого в ней должны быть предусмотрены множества G_S, G_C и O. Для создания реальных БД в ней модель должна поддерживаться СУБД (средний столбец), в которой составляющие модель множества превращаются в соответствующие языковые средства (ЯОД, ЯООЦ и ЯМД). Передавая СУБД команды этих языков, человек создает схему БД (структуры и ограничения целостности), а затем и сами данные (правый столбец). Дерево, приведенное на слайде справа, иллюстрирует взаимосвязи понятий «модель данных» (M), «схема БД» (S_i) и «состояние БД» (D_j).

В последующих трех параграфах мы познакомимся с традиционными взглядами на определение правил структуризации G_S и задания ограничений целостности G_C, а также на операции над данными O. Рассмотрение будем стараться осуществлять максимально обобщенно, тем не менее, иногда будут просматриваться особенности тех или иных моделей. Детальное рассмотрение некоторых из них в следующих главах будем осуществлять по той же схеме – структуры (G_S), ограничения целостности (G_C), операции (O).

Вопросы и задания к параграфу 2.1

1. Объясните, почему человека интересуют не данные, а информация.
2. Что кроме данных необходимо для получения информации?
3. Перечислите и охарактеризуйте три этапа процесса образования информации из данных.
4. Какие разделы семиотики изучают эти этапы?
5. К какому классу информационных систем относятся системы БД?
6. Опишите, как протекает процесс передачи информации о ПрО с использованием систем БД.
7. В чем заключается основное назначение модели данных?
8. Из каких компонентов состоит атомарная единица информации (АЕИ)?
9. Объясните, почему первые три компонента АЕИ являются обязательными.
10. С чем ассоциируются понятия «схема БД» и «база данных» при табличном представлении данных?
11. Дайте структурное определение модели данных.
12. Проиллюстрируйте компоненты модели данных на примере категориальной модели.
13. Что представляет собой СУБД?