Лекция 2

Данные

- Данные — диалектическая составная часть информации. Они представляют собой зарегистрированные сигналы. При этом физический метод регистрации может быть любым: механическое перемещение физических тел, изменение их формы или параметров качества поверхности, изменение электрических, магнитных, оптических характеристик, химического состава и (или) характера химических связей, изменение состояния электронной системы и многое другое. В соответствии с методом регистрации данные могут храниться и транспортироваться на носителях различных видов.

Данные

- Самым распространенным носителем данных является бумага. На бумаге данные регистрируются путем изменения оптических характеристик ее поверхности. Изменение оптических свойств используется также в устройствах, осуществляющих запись лазерным лучом на пластмассовых носителях с отражающим покрытием *(CD-ROM).* В качестве носителей, использующих изменение магнитных свойств, можно назвать магнитные ленты и диски. Регистрация данных путем изменения химического состава поверхностных веществ носителя широко используется в фотографии. На биохимическом уровне происходит накопление и передача данных в живой природе.

Операции с данными

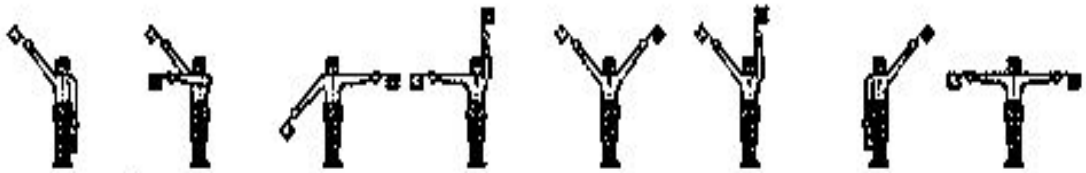
В структуре возможных операций с данными можно выделить следующие основные:

- *сбор данных*—накопление информации с целью обеспечения достаточной полноты для принятия решений;
- *формализация данных* — приведение данных, поступающих из разных источников, к одинаковой форме, чтобы сделать их сопоставимыми между собой, то есть повысить их уровень доступности;
- *фильтрация данных* — отсеивание «лишних» данных, в которых нет необходимости для принятия решений; при этом должен уменьшаться уровень «шума», а достоверность и адекватность данных должны возрастать;
- *сортировка данных* — упорядочение данных по заданному признаку с целью удобства использования; повышает доступность информации;
- *архивация данных* — организация хранения данных в удобной и легкодоступной форме;
- *защита данных* — комплекс мер, направленных на предотвращение утраты, воспроизведения и модификации данных;
- *транспортировка данных*—прием и передача (доставка и поставка) данных между удаленными участниками информационного процесса; при этом источник данных в информатике принято называть *сервером,* а потребителя — *клиентом;*
- *преобразование данных* — перевод данных из одной формы в другую или из одной структуры в другую. Преобразование данных часто связано с изменением типа носителя, например книги можно хранить в обычной бумажной форме, но можно использовать для этого и электронную форму, и микрофотопленку..

Кодирование данных

- Для автоматизации работы с данными, относящимися к различным типам, очень важно унифицировать их форму представления — для этого обычно используется прием *кодирования,* то есть выражение данных одного типа через данные другого типа. Естественные человеческие *языки* — это не что иное, как системы кодирования понятий для выражения мыслей посредством речи. К языкам близко примыкают *азбуки* (системы кодирования компонентов языка с помощью графических символов).

Примеры кодирования данных

C	O	M	P	U	T	E	R	
43	4F	4D	50	55	54	45	52	Код ASCII
								Код Морзе
								Код Брайля
								Код морской сигнальный

Кодирование данных

Своя система существует и в вычислительной технике — она называется *двоичным кодированием* и основана на представлении данных последовательностью всего двух знаков:

0 и 1.

Эти знаки называются *двоичными цифрами,* по-английски — *binary digit или сокращенно bit (бит).*

ГОСТ 8.417-2002 «Единицы величин»:

Русский		Английский (Международный стандарт)	
Полное наименование	Сокращенное наименование	Полное наименование	Сокращенное наименование
бит	бит	bit	bit
байт	Б	byte	B
килобит	Кбит	kilobit	Kbit
килобайт	КБ	kilobyte	KB
мегабит	Мбит	megabit	Mbit
мегабайт	МБ	megabyte	MB
гигабит	Гбит	gigabit	Gbit
гигабайт	ГБ	gigabyte	GB

Кодирование данных

Одним битом могут быть выражены два понятия: 0 или 1 *(да* или *нет, черное* или *белое, истина* или *ложь* и т. п.). Если количество битов увеличить до двух, то уже можно выразить четыре различных понятия: 00 01 10 11

Тремя битами можно закодировать восемь различных значений: 000 001 010 011 100 101 110 111

Увеличивая на единицу количество разрядов в системе двоичного кодирования, мы увеличиваем в два раза количество значений, которое может быть выражено в данной системе, то есть общая формула имеет вид:

$$N=2^m,$$

- где N— количество независимых кодируемых значений;
- m — разрядность двоичного кодирования, принятая в данной системе.

Кодирование целых чисел

- Целые числа кодируются двоичным кодом достаточно просто — достаточно взять целое число и делить его пополам до тех пор, пока частное не будет равно единице. Совокупность остатков от каждого деления, записанная справа налево вместе с последним частным, и образует двоичный аналог десятичного числа.
- 19:2 = 9+1
- 9:2 = 4 + 1
- 4:2=2+0
- 2:2=1+0
- Таким образом, $19_{10} = 10011_{2}$.

Кодирование целых чисел

- Для кодирования целых чисел от 0 до 255 достаточно иметь 8 разрядов двоичного кода (8 бит). Шестнадцать бит позволяют закодировать целые числа от 0 до 65 535, а 24 бита — уже более 16,5 миллионов разных значений.
- Для кодирования действительных чисел используют 80-разрядное кодирование. При этом число предварительно преобразуется в *нормализованную форму:*
- $3{,}1415926 = 0{,}31415926 \cdot 10^{1}$
- $300\,000 = 0{,}3 \cdot 10^{6}$
- $123\,456\,789 = 0{,}123456789 \cdot 10^{10}$
- Первая часть числа называется *мантиссой,* а вторая *— характеристикой.*

Кодирование текстовых данных

- Если каждому символу алфавита сопоставить определенное целое число (например, порядковый номер), то с помощью двоичного кода можно кодировать и текстовую информацию. Восьми двоичных разрядов достаточно для кодирования 256 различных символов. Этого хватит, чтобы выразить различными комбинациями восьми битов все символы английского и русского языков, как строчные, так и прописные, а также знаки препинания, символы основных арифметических действий и некоторые общепринятые специальные символы, например символ «§».

Кодирование текстовых данных

- Первые 32 кода базовой таблицы, начиная с нулевого, отданы производителям аппаратных средств (в первую очередь производителям компьютеров и печатающих устройств). В этой области размещаются так называемые *управляющие коды,* которым не соответствуют никакие символы языков, и, соответственно, эти коды не выводятся ни на экран, ни на устройства печати, но ими можно управлять тем, как производится вывод прочих данных.
- Начиная с кода 32 по код 127 размещены коды символов английского алфавита, знаков препинания, цифр, арифметических действий и некоторых вспомогательных символов.

Кодирование текстовых данных

Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	Null	32	20	Space	64	40	@	96	60	`
1	01	Start of heading	33	21	!	65	41	A	97	61	a
2	02	Start of text	34	22	"	66	42	B	98	62	b
3	03	End of text	35	23	#	67	43	C	99	63	c
4	04	End of transmit	36	24	$	68	44	D	100	64	d
5	05	Enquiry	37	25	%	69	45	E	101	65	e
6	06	Acknowledge	38	26	&	70	46	F	102	66	f
7	07	Audible bell	39	27	'	71	47	G	103	67	g
8	08	Backspace	40	28	(	72	48	H	104	68	h
9	09	Horizontal tab	41	29	)	73	49	I	105	69	i
10	0A	Line feed	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage return	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	47	2F	/	79	4F	O	111	6F	o
16	10	Data link escape	48	30	0	80	50	P	112	70	p
17	11	Device control 1	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	50	32	2	82	52	R	114	72	r
19	13	Device control 3	51	33	3	83	53	S	115	73	s
20	14	Device control 4	52	34	4	84	54	T	116	74	t
21	15	Neg. acknowledge	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	54	36	6	86	56	V	118	76	v
23	17	End trans. block	55	37	7	87	57	W	119	77	w
24	18	Cancel	56	38	8	88	58	X	120	78	x
25	19	End of medium	57	39	9	89	59	Y	121	79	y
26	1A	Substitution	58	3A	:	90	5A	Z	122	7A	z
27	1B	Escape	59	3B	;	91	5B	[	123	7B	{
28	1C	File separator	60	3C	<	92	5C	\	124	7C	|
29	1D	Group separator	61	3D	=	93	5D	]	125	7D	}
30	1E	Record separator	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	63	3F	?	95	5F	_	127	7F	□

Кодирование текстовых данных

- Аналогичные системы кодирования текстовых данных были разработаны и в других странах. Так, например, в СССР в этой области действовала система кодирования КОИ-7 *(код обмена информацией, семизначный).* Однако поддержка производителей оборудования и программ вывела американский код *ASCII* на уровень международного стандарта, и национальным системам кодирования пришлось «отступить» во вторую, расширенную часть системы кодирования, определяющую значения кодов со 128 по 255. Отсутствие единого стандарта в этой области привело к множественности одновременно действующих кодировок. Только в России можно указать три действующих стандарта кодировки и еще два устаревших.

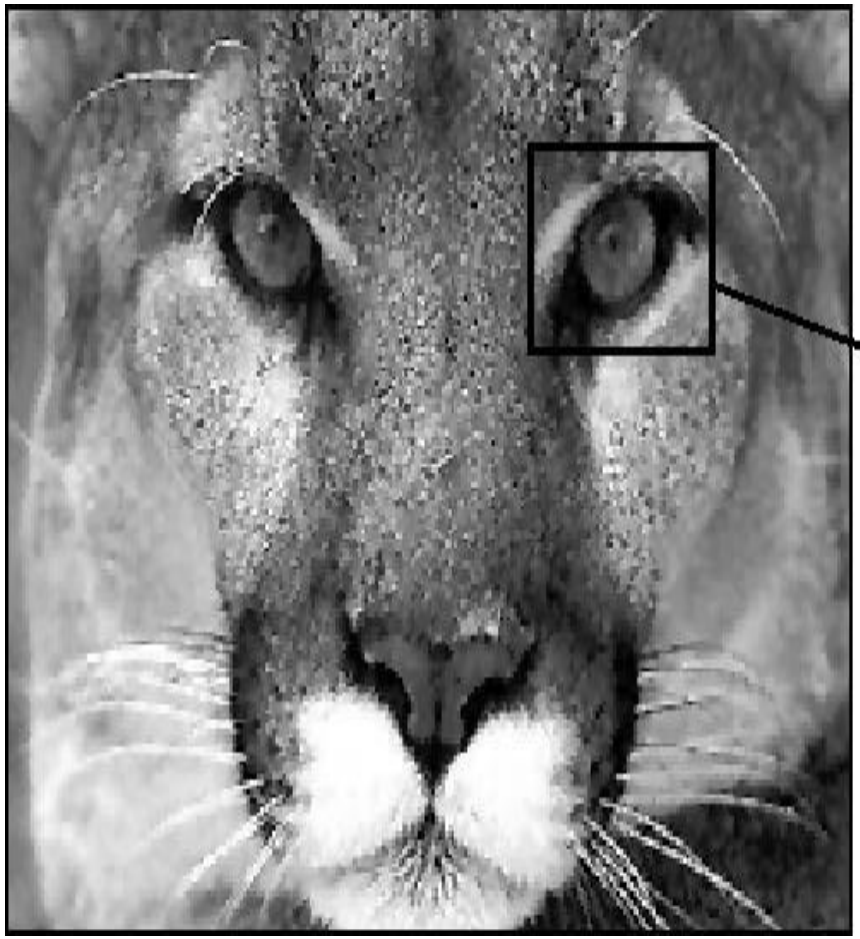
Кодирование текстовых данных

- Так, например, кодировка символов русского языка, известная как кодировка *Windows-1251,* была введена «извне» — компанией Microsoft, но, учитывая широкое распространение операционных систем и других продуктов этой компании в России, она глубоко закрепилась и нашла широкое распространение (таблица 1.2). Эта кодировка используется на большинстве локальных компьютеров, работающих на платформе Windows.

Кодирование текстовых данных

- Другая распространенная кодировка носит название КОИ-8 *(код обмена информацией, восьмизначный)* — ее происхождение относится ко временам действия Совета Экономической Взаимопомощи государств Восточной Европы (таблица 1.3). Сегодня кодировка КОИ-8 имеет широкое распространение в компьютерных сетях на территории России и в российском секторе Интернета.

Кодирование графических данных



Растр — это метод кодирования графической информации, издавна принятый в полиграфии

Кодирование графических данных

- Для кодирования цветных графических изображений применяется *принцип декомпозиции* произвольного цвета на основные составляющие. В качестве таких составляющих используют три основные цвета: красный *(Red, К),* зеленый *(Green,* G) и синий *(Blue, B).* На практике считается (хотя теоретически это не совсем так), что любой цвет, видимый человеческим глазом, можно получить путем механического смешения этих трех основных цветов. Такая система кодирования называется системой *RGB* по первым буквам названий основных цветов.

Кодирование графических данных

- Если для кодирования яркости каждой из основных составляющих использовать по 256 значений (восемь двоичных разрядов), как это принято для полутоновых черно-белых изображений, то на кодирование цвета одной точки надо затратить 24 разряда. При этом система кодирования обеспечивает однозначное определение 16,5 млн различных цветов, что на самом деле близко к чувствительности человеческого глаза. Режим представления цветной графики с использованием 24 двоичных разрядов

Кодирование звуковой информации

Приемы и методы работы со звуковой информацией пришли в вычислительную технику наиболее поздно. К тому же, в отличие от числовых, текстовых и графических данных, у звукозаписей не было столь же длительной и проверенной истории кодирования. В итоге методы кодирования звуковой информации двоичным кодом далеки от стандартизации. Множество отдельных компаний разработали свои корпоративные стандарты, но если говорить обобщенно, то можно выделить два основных направления.

Кодирование звуковой информации

Приемы и методы работы со звуковой информацией пришли в вычислительную технику наиболее поздно. К тому же, в отличие от числовых, текстовых и графических данных, у звукозаписей не было столь же длительной и проверенной истории кодирования. В итоге методы кодирования звуковой информации двоичным кодом далеки от стандартизации. Множество отдельных компаний разработали свои корпоративные стандарты, но если говорить обобщенно, то можно выделить два основных направления.

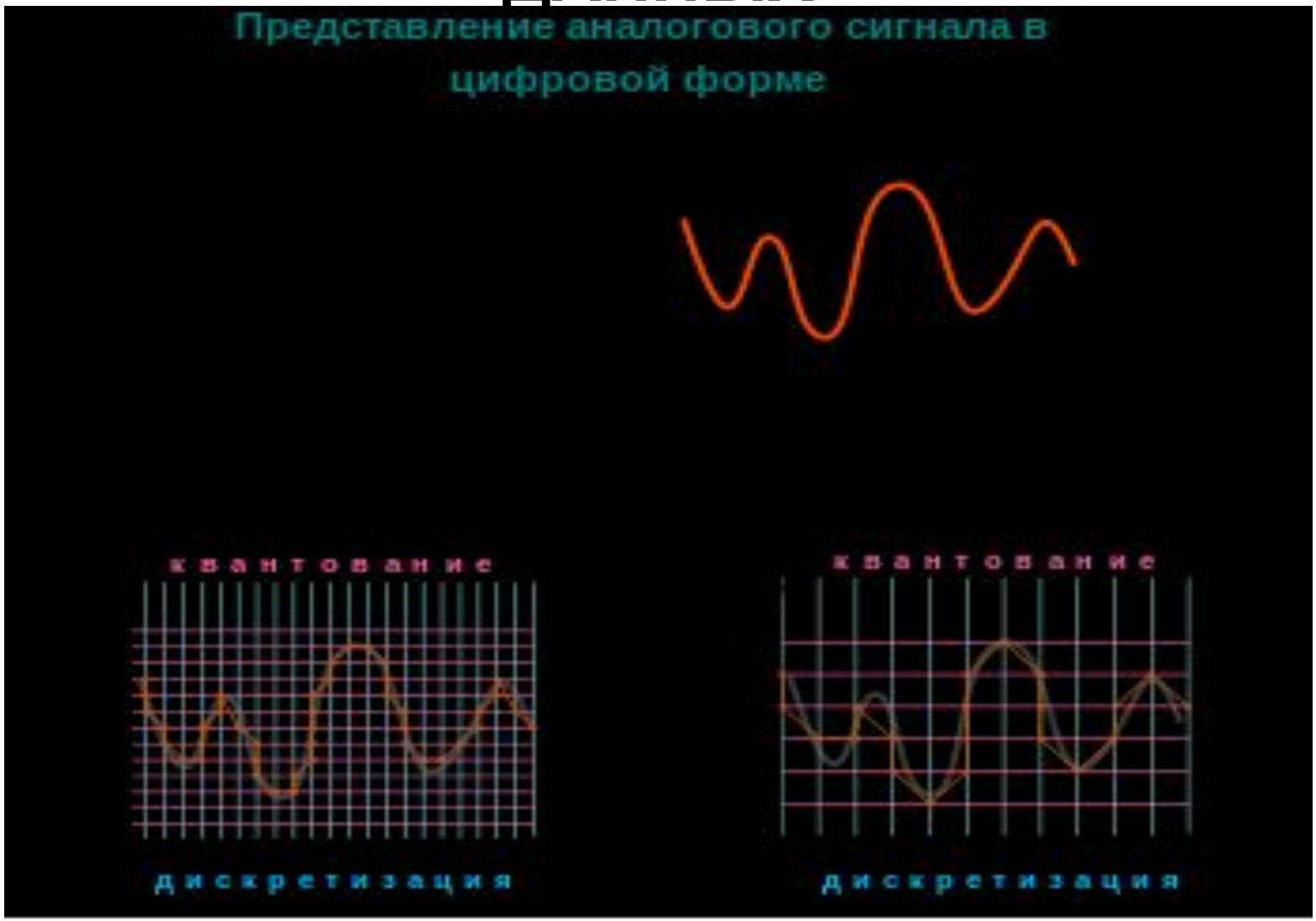
Кодирование звуковой информации

- Метод FM *(Frequency Modulation)* основан на том, что теоретически любой сложный звук можно разложить на последовательность простейших гармонических сигналов разных частот, каждый из которых представляет собой правильную синусоиду, а следовательно, может быть описан числовыми параметрами, то есть кодом. В природе звуковые сигналы имеют непрерывный спектр, то есть являются аналоговыми.

Кодирование звуковой информации

Их разложение в гармонические ряды и представление в виде дискретных цифровых сигналов выполняют специальные устройства — *аналогово-цифровые преобразователи (АЦП).* Обратное преобразование для воспроизведения звука, закодированного числовым кодом, выполняют *цифро-аналоговые преобразователи* (ДАЛ). При таких преобразованиях неизбежны потери информации, связанные с методом кодирования, поэтому качество звукозаписи обычно получается не вполне удовлетворительным и соответствует качеству звучания простейших электромузыкальных инструментов с окрасом, характерным для

Кодирование графических данных



Кодирование звуковой информации

- Метод таблично-волнового (*Wave-Table)* синтеза лучше соответствует современному уровню развития техники. Если говорить упрощенно, то можно сказать, что где-то в заранее подготовленных таблицах хранятся образцы звуков для множества различных музыкальных инструментов (хотя не только для них). В технике такие образцы называют *сэмплами.* Числовые коды выражают тип инструмента, номер его модели, высоту тона, продолжительность и интенсивность звука, динамику его изменения, некоторые параметры среды, в которой происходит звучание, а также прочие параметры, характеризующие особенности звука. Поскольку в качестве образцов используются «реальные» звуки, то качество звука, полученного в результате синтеза, получается очень высоким и приближается к качеству звучания реальных музыкальных инструментов.

Основные структуры данных

Работа с большими наборами данных автоматизируется проще, когда данные *упорядочены,* то есть образуют заданную структуру. Существует три основных типа структур данных: *линейная, иерархическая* и *табличная.*

Рассмотрим на примере обычной книги.

Основные структуры данных. Линейная

- Если разобрать книгу на отдельные листы и перемешать их, книга потеряет свое назначение. Она по-прежнему будет представлять набор данных, но подобрать адекватный метод для получения из нее информации весьма непросто. (Еще хуже дело будет обстоять, если из книги вырезать каждую букву отдельно — в этом случае вряд ли вообще найдется адекватный метод для ее прочтения.)
- Если же собрать все листы книги в правильной последовательности, мы получим простейшую структуру данных — *линейную.* Такую книгу уже можно читать, хотя для поиска нужных данных ее придется прочитать подряд, начиная с самого начала, что не всегда удобно.

Основные структуры данных. Линейная

- Тогда нужный элемент можно разыскать по номеру строки.
- N п/п Фамилия, Имя, Отчество
- 1Аистов Александр Алексеевич
- 2Бобров Борис Борисович
- 3Воробьева Валентина Владиславовна
- ……..
- 27 Сорокин Сергей Семенович
- Разделителем может быть и какой-нибудь специальный символ. Нам хорошо известны разделители между словами — это пробелы. В русском и во многих европейских языках общепринятым разделителем предложений является точка. В рассмотренном нами классном журнале в качестве разделителя можно использовать любой символ, который не встречается в самих данных, например символ «*». Тогда наш список выглядел бы так:
- Аистов Александр Алексеевич * Бобров Борис Борисович * Воробьева Валентина Владиславовна *... * Сорокин Сергей Семенович
- В этом случае для розыска элемента с номером *n* надо просмотреть список начиная с самого начала и пересчитать встретившиеся разделители. Когда будет отсчитано *n-i* разделителей, начнется нужный элемент. Он закончится, когда будет встречен следующий разделитель.

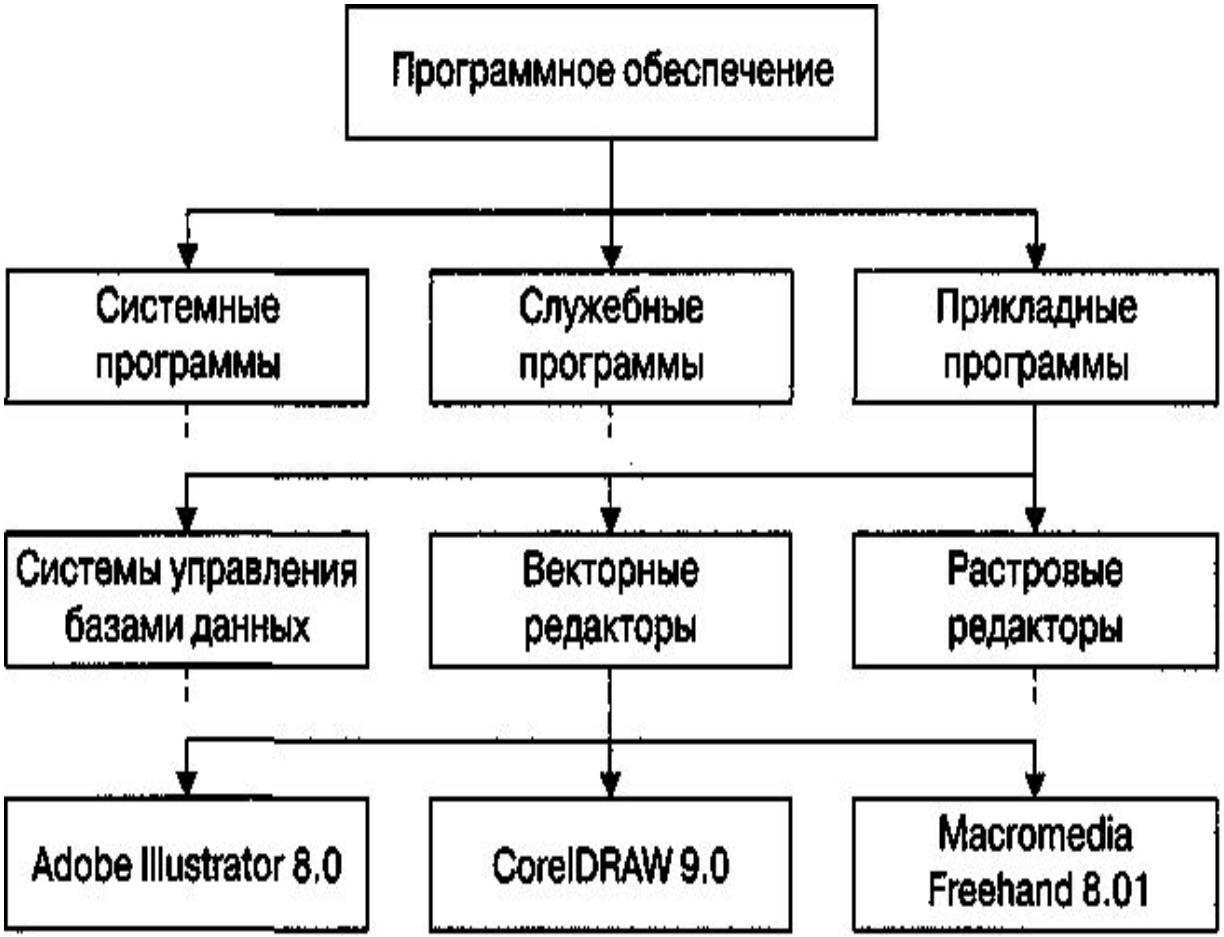
Основные структуры данных. Линейная

- Таким образом, *линейные структуры данных (списки) — это упорядоченные структуры, в которых адрес элемента однозначно определяется его номером.*

Основные структуры данных. Иерархическая.

Для быстрого поиска данных существует *иерархическая структура.* Так, например, книги разбивают на части, разделы, главы, параграфы и т. п. Элементы структуры более низкого уровня входят в элементы структуры более высокого уровня: разделы состоят из глав, главы из параграфов и т. д.

Основные структуры данных. Иерархическая.



Основные структуры данных. Табличная.

На практике задачу упрощают тем, что в большинстве книг есть вспомогательная перекрестная *таблица,* связывающая элементы иерархической структуры с элементами линейной структуры, то есть связывающая разделы, главы и параграфы с номерами страниц. В книгах с простой иерархической структурой, рассчитанных на последовательное чтение, эту таблицу принято называть *оглавлением,* а в книгах со сложной структурой, допускающей выборочное чтение, ее называют

Основные структуры данных. Табличная.

Планета	Расстояние до Солнца, а.е.	Относительная масса	Количество спутников
Меркурий	0,39	0,056	0
Венера	0,67	0,88	0
Земля	1,0	1,0	1
Марс	1,51	0,1	2
Юпитер	5,2	318	16

Основные структуры данных. Табличная.

Если нужно сохранить таблицу в виде длинной символьной строки, используют один символ-разделитель между элементами, принадлежащими одной строке, и другой разделитель для отделения строк, например так:

Меркурий*0,39*0,056*0#Ввнера*0,67*0,88*0#Земля*1,0*1(0*1#Марс*1)б1*0,1*2#..

Единицы измерения данных

- В информатике для измерения данных используют тот факт, что разные типы данных имеют универсальное двоичное представление, и потому вводят свои единицы данных, основанные на нем.
- Наименьшей единицей измерения является байт.

Единицы измерения данных

- Более крупная единица измерения — килобайт (Кбайт).
- 1 Кбайт равен 2^{10} байт (1024 байт)

Более крупные единицы измерения данных образуются добавлением префиксов *мега-, гига-, тера-*

- 1 Мбайт = 1024 Кбайт = 2^{20} байт
- 1 Гбайт = 1024 Мбайт = 2^{30} байт
- 1 Тбайт = 1024 Гбайт = 2^{40} байт

Единицы хранения данных

- В качестве единицы хранения данных принят объект переменной длины, называемый *файлом. Файл — это последовательность произвольного числа байтов, обладающая уникальным собственным именем.* Обычно в отдельном файле хранят данные, относящиеся к одному типу. В этом случае тип данных определяет *тип файла.*

Понятие о файловой структуре

- Хранение файлов организуется в иерархической структуре, которая в данном случае называется *файловой структурой.* В качестве вершины структуры служит имя носителя, на котором сохраняются файлы. Далее файлы группируются в *каталоги (папки),* внутри которых могут быть созданы *вложенные каталоги (папки). Путь доступа к файлу* начинается с имени устройства и включает все имена каталогов (папок), через которые проходит. В качестве разделителя используется символ «\» (обратная косая черта).

Понятие о файловой структуре

Пример записи полного имени файла:

<имя носителя>\<имя каталога-1>\...\<имя каталога-М>\<собственное имя файла