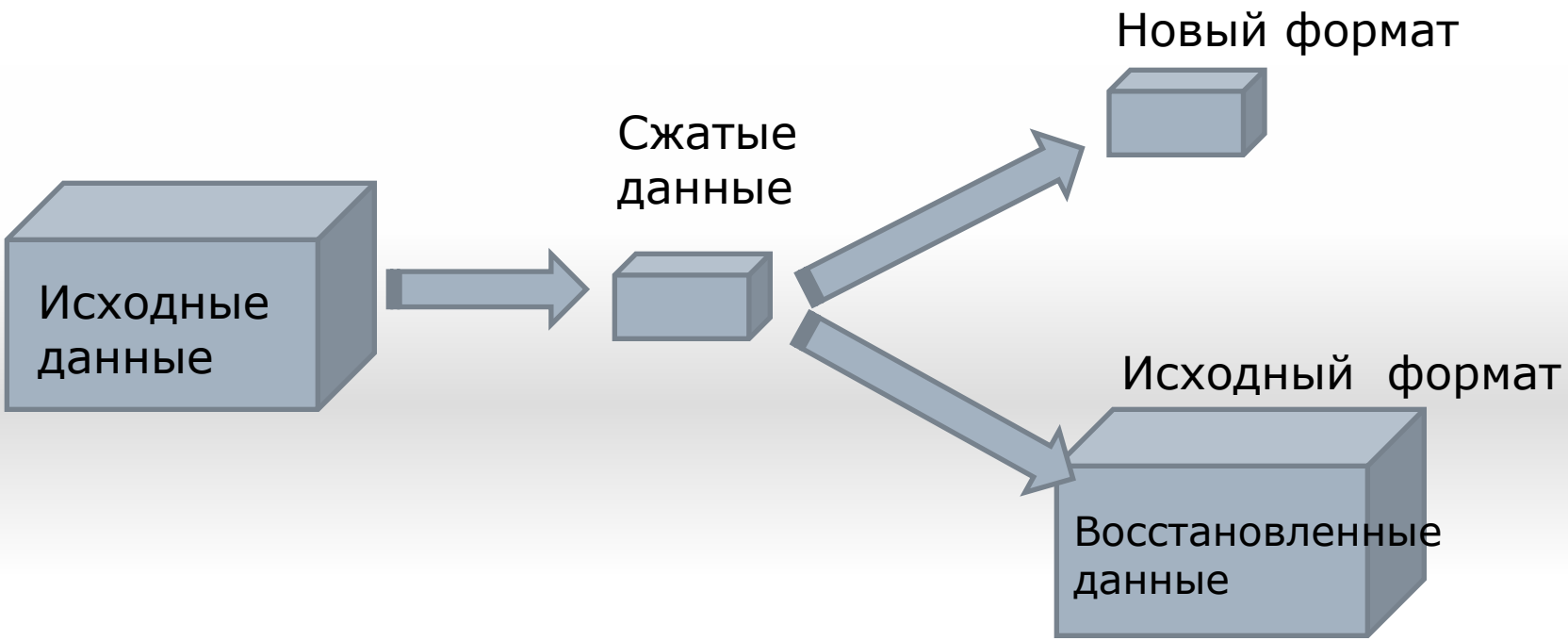
Сжатие информации Архивация файлов

Сжатие данных это процесс, обеспечивающий уменьшение объема данных.

Способы сжатия

- Изменение содержания данных (уменьшение избыточности данных)
- Изменение структуры данных (эффективное кодирование)
- Изменение содержания и структуры данных

Цели сжатия данных – экономия ресурсов при хранении или передаче данных



Архивация данных – сжатие с возможностью полного восстановления данных

Архивация – процесс, при котором файлы сжимаются без потери информации.

При **разархивации** данные и программы восстанавливаются в исходном виде.

Архиватор – программа, выполняющая сжатие (архивирование) файлов для более компактного хранения во внешней памяти и восстановление (разархивирование) сжатых файлов в первоначальном состоянии.

Коэффициент сжатия – это величина для обозначения эффективности метода сжатия, равная отношению количества информации до и после сжатия.

$$К_{сж} = 2 \text{ МБ} / 0{,}5 \text{ МБ} = 4$$

Размер файла 2МБ

Размер файла 512 КБ

Сжатие данных может происходить с потерями и без потерь

- **Сжатие без потерь (полностью обратимое)** – это методы сжатия данных, при которых данные восстанавливаются после их распаковки полностью без внесения изменений (используется для текстов, программ) Ксж до 50%
- **Сжатие с регулируемыми потерями** – это методы сжатия данных, при которых часть данных отбрасывается и не подлежит восстановлению (используется для видео, звука, изображений) Ксж до 99%

Сжатие с потерями

Тип данных	Тип файла после сжатия	Степень сжатия
Графика	.JPG	до 99%
Видео	.MPG	
Звук	.MP3	

Сжатие без потерь

Тип данных	Тип файла после сжатия	Степень сжатия
Графика	.GIF .TIF .PCX	До 50%
Видео	.AVI	
Любой тип	.ZIP .ARJ .RAR .LZH	

Алгоритмы сжатия символьных данных

- **Статистические методы** – это методы сжатия, основанные на статистической обработке текста.
- **Словарное сжатие** – это методы сжатия, основанные на построении внутреннего словаря.

Алгоритмы и методы архивации

1. Алгоритм сжатия данных (RLE) основан на замене повторяющихся битов (в тексте может иметься последовательность одинаковых символов, в графическом файле – закрашенная одним цветом область и т.д.). При применении этого алгоритма вместо последовательности одинаковых по цвету пикселей в строке изображения записывается цвет и количество его повторений. Такой подход используется при хранении изображений в формате BMP.

Алгоритмы и методы архивации

Например, в тексте подряд идут 10 пробелов, которые кодируются 10-тью байтами. При архивации они заменяются 3-мя байтами (первый байт – кодирует заменяемый символ; второй байт – специальный байт "флажка" архивации, который указывает на необходимость развернуть первый байт в последовательность байтов; третий байт указывает количество повторяющихся байтов).

Алгоритмы и методы архивации

2. Алгоритм кодирования одинаковых последовательностей символов (LZW, назван по фамилиям авторов Якоб Лемпель, Абрахам Зив и Терри Велч) кодирует повторяющиеся фрагменты (слова, "узоры") определенным кодом (последовательностью бит) и при их повторном появлении заменяет ссылкой на первичный код, хранящийся в специально создаваемой таблице (словаре).

Алгоритм разработан израильскими математиками Якобом Зивом и Аб рахам ом Лемпелем.

Словарь содержит, кроме многих других, такие цепочки:
1-ра 2-аб 3-ат 4-мат 5-ми_ 6-ам 7-бо
8-ом_ 9-бом 10-ем 11-лем

Алгоритм раз1ботан из1ильскими мате4ика5Яко7ом Зив821х68 Л10пе11

Чем длиннее цепочка,заменяемая ссылкой в словарь, тем больше эффект сжатия.

Сжатие с потерями

С появлением средств оцифровки изображений появилась существенная проблема: в фотоизображениях практически не встречались точно повторяющиеся последовательности точек. С учетом роста частоты дискретизации и небольшой емкости носителей, это затрудняло их обработку и применение. Фактически средний жесткий диск мог хранить только 45–50 изображений высокого качества.

Сжатие с потерями

Для решения этой проблемы группой специалистов был разработан специальный формат и способ сжатия, получивши название **JPEG** (***J****oint* ***P****hotographic* ***E****xpert* ***G****roup*, объединенная группа экспертов-фотографов). Алгоритм сжатия, предложенный ими, подразумевал **сжатие с потерей качества**. Его достоинством было то, что «силу» сжатия можно было указывать изначально и таким образом находить компромисс между качеством и объемом изображения. Первый стандарт этого алгоритма был принят в 1991 году.

Сжатие с потерями

Следующим шагом стала разработка группы методов, предназначенных для сжатия потоковых данных (видео и аудио). Существенной особенностью этих данных является их очень большой объем и постепенное изменение (из-за высокой частоты между двумя соседними кадрами, как правило, разница невелика). Сжатый видео- и/или аудиопоток характеризуется чаще всего общим показателем **битрейтом** (*bit rate* — битовая скорость) — количеством битов на одну секунду использования, которое получается после упаковки.

Сжатие с потерями

Первым был разработан и принят в 1992 году стандарт MPEG-1, включавший в себя способ сжатия видео в поток до 1,5 Мбит, аудио в поток 64, 128 или 192 Кбит/с на канал, а также алгоритмы синхронизации. Стандарт описывал не алгоритмы, а формат получающегося битового потока. Это позволило в дальнейшем разработать множество реализаций алгоритмов кодирования и декодирования. Стандарт применялся для создания видео и CD.

Сжатие с потерями

Особенную популярность завоевала реализация MPEG-1 для упаковки звука. Применяется для этого стандарт **MPEG-1 Layer 3** (сокращенно названный **MP3**). При сжатии этим методом используется сжатие с потерей информации. Причем учитывается особенность слухового восприятия: если рядом расположены две частоты, то более громкая «перекрывает» более тихую. Таким образом, ее можно сгладить без ощутимой потери качества звука.

Сжатие с потерями

Следующим шагом была разработка и принятие в 1995 году стандарта MPEG-2, предусматривающего работу с более качественным видеопотоком, скорость которого могла изменяться от 3 до 10 Мбит/с. Эта группа методов применяется при создании DVD-дисков.

Группа стандартов, получившая позднее название **MPEG-4**, изначально проектировалась для работы с очень низкими потоками, но в дальнейшем претерпела много изменений. В основном эти изменения касались введения интеллектуальных методов — например, описания параметров отображения лиц или синтеза речи.

Существуют различные методы архивации файлов (ZIP, RAR, ARJ и др.), которые используют различные алгоритмы архивации и различаются степенью сжатия файлов, скоростью выполнения и другими параметрами.

Лучше всего сжимаются текстовые и графические файлы и практически не сжимаются файлы архивов.

Параметры архивации

1. *Многотомные архивы*, т.е. архивы, состоящие из нескольких частей (используются для сохранения большого архива на нескольких дисках). Первый том архива имеет обычное расширение rar, а расширения последующих томов нумеруются как r00, r01, r02 и т.д.

2. *Непрерывные архивы* (максимальная степень сжатия).

Параметры архивации

3. *Самораспаковывающиеся* архивы (SFX, от англ. SelF-eXtracting). Для разархивации такого архива не нужна специальная программа, достаточно запустить файл архива на выполнение, т.к. он является исполняемым файлом и имеет расширение exe.

4. Архивы, созданные с использованием метода *мультимедиа-сжатие* (на 30% более высокая степень сжатия, чем при обычном).

Методы архивации:

- Без сжатия (просто помещает файлы в архив без их упаковки).
- Скоростной (сжимает плохо, но очень быстро).
- Быстрый.
- Обычный (используется для создания резервных копий данных).
- Хороший.
- Максимальный (обеспечивает наиболее высокую степень сжатия, но с наименьшей скоростью, используется для передачи по компьютерным сетям или для долговременного хранения).

При создании нового архива нужно задать:

- имя архивного файла и его место хранения на диске;
- формат архивации (RAR, ZIP или др.);
- параметры архивирования.