Сжатие информации

План.

1. Понятие о сжатии информации.
2. Программы архиваторы.

Избыточность

Редакторы, работающие с текстовой, графической, звуковой и другой информацией, **кодируют** ее наиболее **естественным**, но **не самым экономичным способом**.

Действительно, если внимательно посмотреть любой текст, то можно заметить, что такие буквы «а» и «о», встречаются в нем гораздо чаще чем «ю» и «у». То же самое можно отнести и к сочетаниям букв.

На рисунках цвета соседних точек в большинстве случаев близки по оттенку. Подобно этому в любой последовательности информации некоторые сочетания встречаются намного чаще других.

Избыточность

Все это приводит к тому, что в файлах, хранящих эту информацию, некоторые **комбинации** из 0 и 1 **встречаются гораздо чаще**, чем другие. В таких случаях говорят, что информация обладает **избыточностью**, и есть возможность перекодировать содержание файла, уменьшив его размер.

Для **сжатия** достаточно придерживаться правила: чем чаще встречается комбинация, тем более коротким сочетанием из 0 и 1 ее можно **перекодировать**. Разумеется, делать это должна программа.

Сжатие данных

– это процесс, обеспечивающий уменьшение объема данных путем сокращения их избыточности. Сжатие данных связано с компактным расположением порций данных стандартного размера.

Сжатие происходит за счет устранения избыточности кода, например, за счет упрощения кодов, исключения из них постоянных битов или представления повторяющихся символов в виде коэффициента повторения.

1. Равномерное сжатие с использованием кодов одной длины.

Этот метод используется, если в записи сообщения присутствует небольшая часть алфавита.

2. Сжатие с использованием кодов переменной длины.

Сокращение объёма данных достигается за счёт замены часто встречающихся данных короткими кодовыми словами, а редких — длинными .

Сжатие данных можно разделить на два основных типа:

Сжатие без потерь (полностью обратимое) – это метод сжатия данных, при котором ранее закодированная порция данных восстанавливается после их распаковки полностью без внесения изменений. Для каждого типа данных, как правило, существуют свои оптимальные алгоритмы сжатия без потерь.

Сжатие с потерями – это метод сжатия данных, при котором для обеспечения *максимальной степени* сжатия исходного массива данных часть содержащихся в нем данных отбрасывается. Для текстовых, числовых и табличных данных использование программ, реализующих подобные методы сжатия, является неприемлемыми. В основном такие алгоритмы

ОБРАТИМОСТЬ СЖАТИЯ.

Характерными форматами сжатия с потерей информации являются:

.JPG для графических данных;

.MPG для видеоданных;

.MP3 для звуковых данных.

Характерными форматами сжатия без потери информации являются:

.GIF,.TIF,.PCX и многие другие для графических данных;

.AVI для видеоданных;

.ZIP,.ARJ,.RAR,.LZH,.LH,.CAB и многие другие для любых типов данных

Алгоритм сжатия данных (алгоритм архивации) – это *алгоритм*, который устраняет *избыточность* записи данных.

Алфавит кода – множество всех символов входного потока.

При сжатии англоязычных текстов обычно используют множество из 128 *ASCII* кодов.

При сжатии изображений множество значений пиксела может содержать 2, 16, 256 или другое количество элементов.

Кодовый символ – наименьшая *единица* данных, подлежащая сжатию. Обычно символ – это 1 *байт*, но он может быть битом, тритом {0,1,2}, или чем-либо еще.

Кодовое слово – это последовательность кодовых символов из алфавита кода.

Если все слова имеют одинаковую длину (число символов), то такой код называется *равномерным (фиксированной длины)*, а если же допускаются слова разной длины, то – *неравномерным (переменной длины)*.

Код – полное множество слов.

Токен – *единица* данных, записываемая в сжатый *поток* некоторым алгоритмом сжатия. *Токен* состоит из нескольких полей фиксированной или переменной длины.

Фраза – фрагмент данных, помещаемый в словарь для дальнейшего использования в сжатии.

Кодирование – процесс сжатия данных.

Декодирование – *обратный* кодированию процесс, при котором осуществляется восстановление данных.

Отношение сжатия – одна из наиболее часто используемых величин для обозначения эффективности метода сжатия.

$$Отношение\ сжатия = \frac{размер\ выходного\ потока}{размер\ входного\ потока}$$

Коэффициент сжатия – величина, обратная отношению сжатия.

$$Коэффициент\ сжатия = \frac{размер\ входного\ потока}{размер\ выходного\ потока}$$

Средняя длина кодового слова – это величина, которая вычисляется как взвешенная вероятностями сумма длин всех кодовых слов.

$$L_{ср}=p_1L_1+p_2L_2+...+p_nL_n,$$

где – вероятности кодовых слов;

$L_1,L_2,...,L_n$ – длины кодовых слов.

Существуют два основных способа проведения сжатия.

Статистические методы – методы сжатия, присваивающие коды переменной длины символам входного потока, причем более короткие коды присваиваются символам или группам символам, имеющим большую *вероятность* появления во входном потоке. Лучшие *статистические методы* применяют *кодирование* Хаффмана.

Словарное сжатие – это методы сжатия, хранящие фрагменты данных в "словаре" (некоторая *структура данных*). Если строка новых данных, поступающих на вход, идентична какому-либо фрагменту, уже

3.Архиваторы

- Программы, осуществляющие сжатие (упаковку файлов), называют архиваторами.

При сжатии можно уменьшить размер файла в несколько раз, что дает заметную экономию памяти.

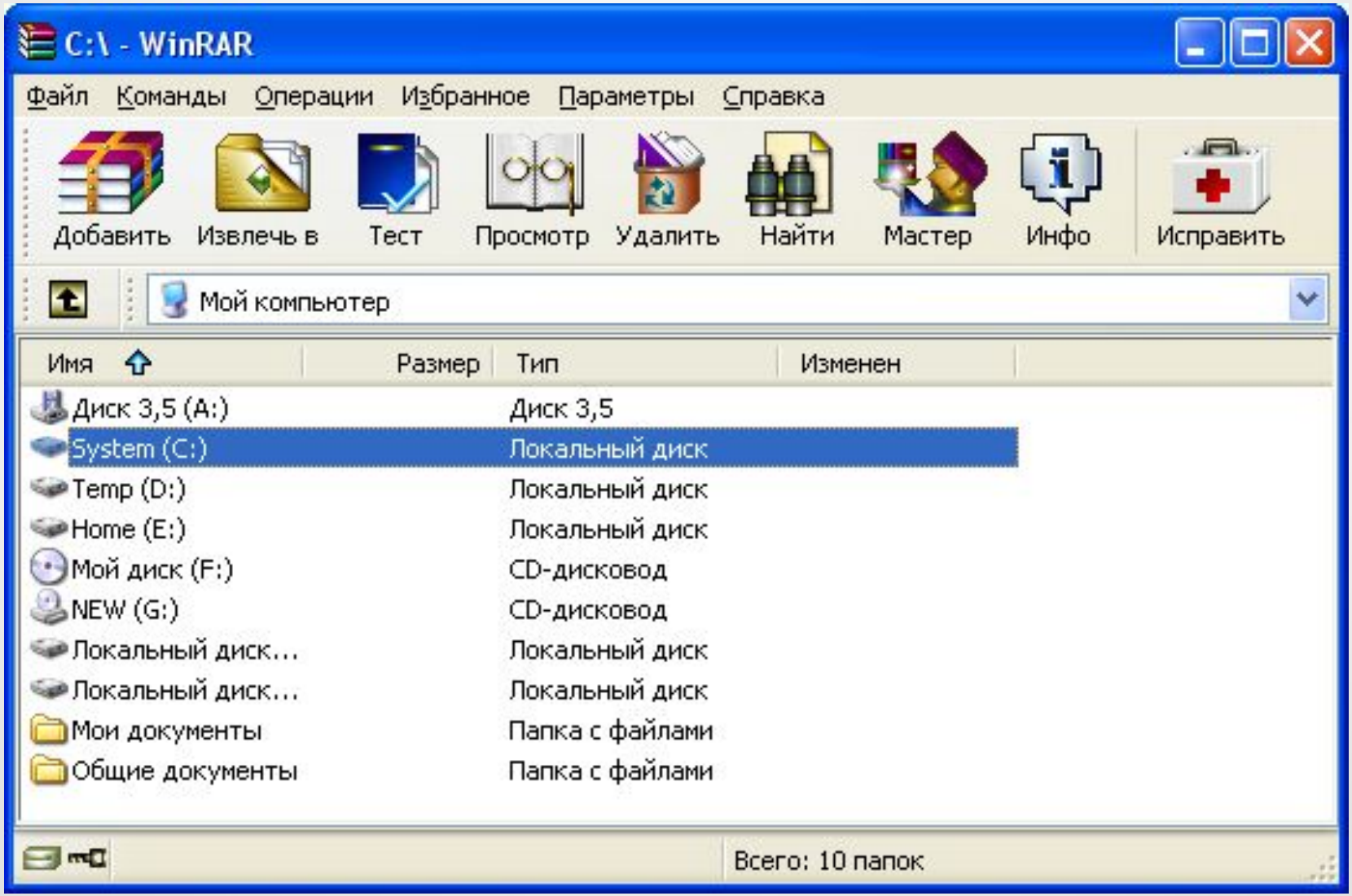
Например: WinRar и WinZip

Пуск→Все программы→Архиваторы→ WinRAR→ WinRAR

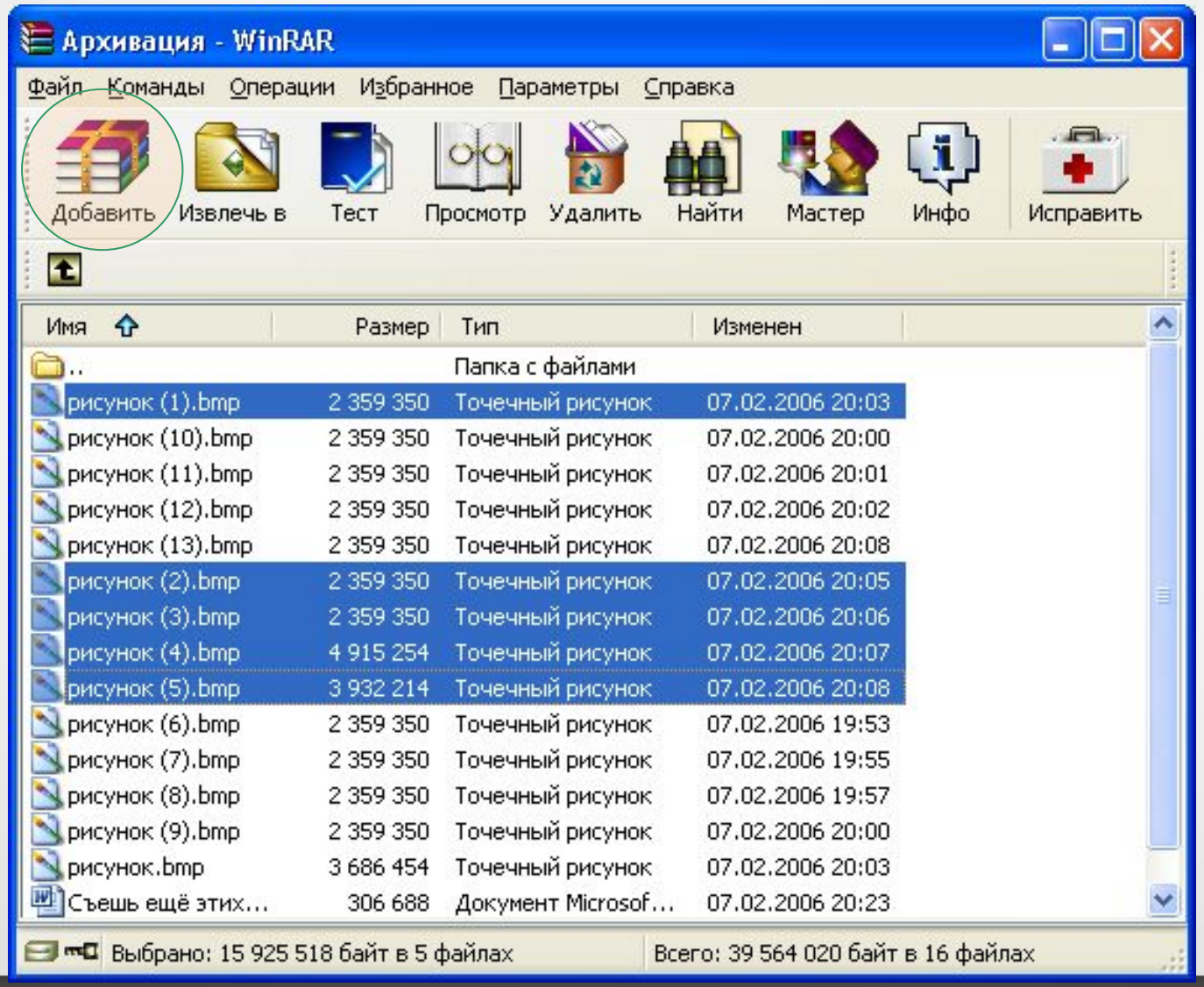
Основные действия при работе с архивами:

- создание нового архива;
- добавление файлов в архив;
- просмотр содержимого архива;
- извлечение файлов из архива;
- просмотр файла в архиве;
- удаление файлов из архива.

Оболочка WinRAR



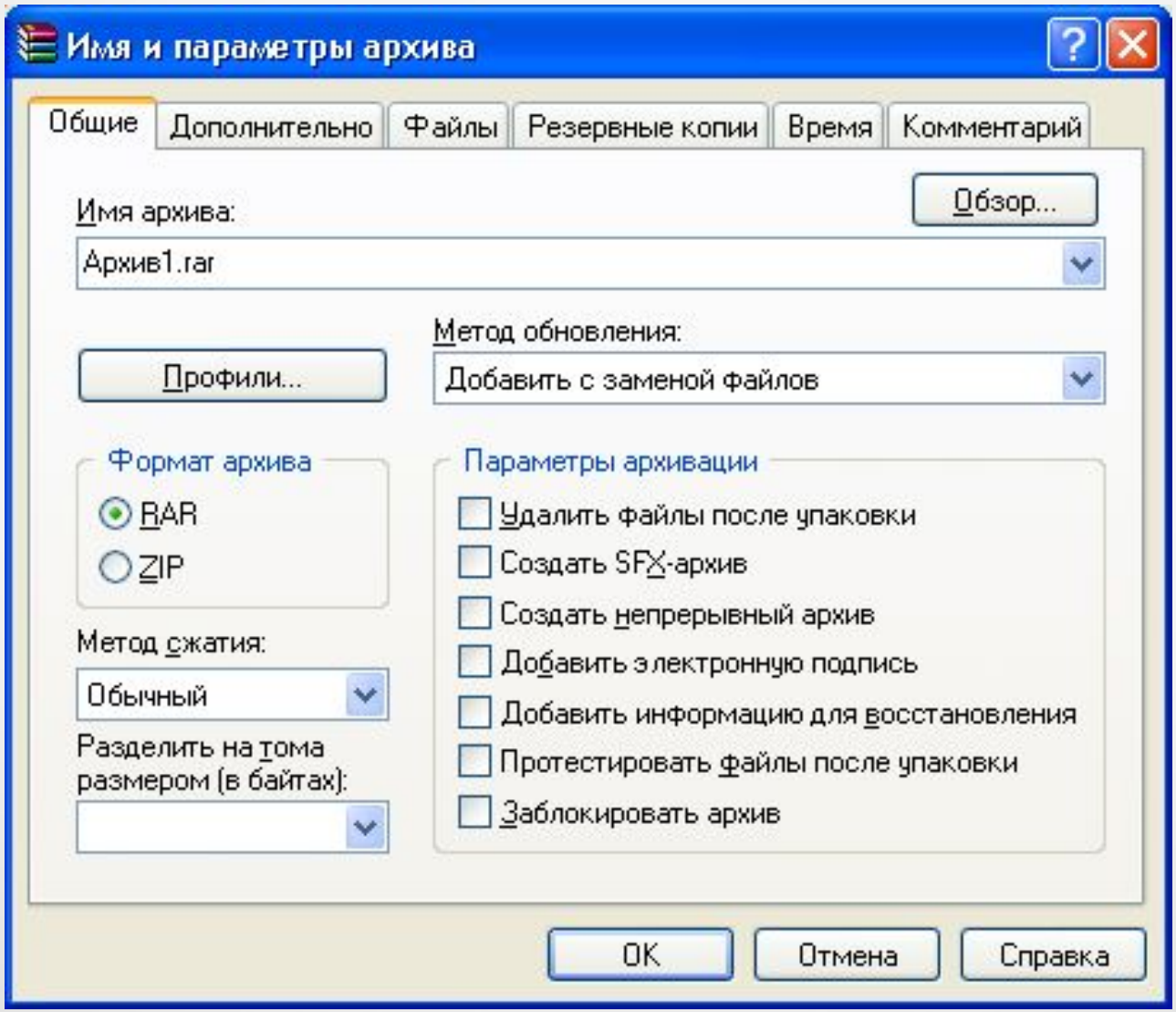
Архивация с помощью оболочки WinRAR



1

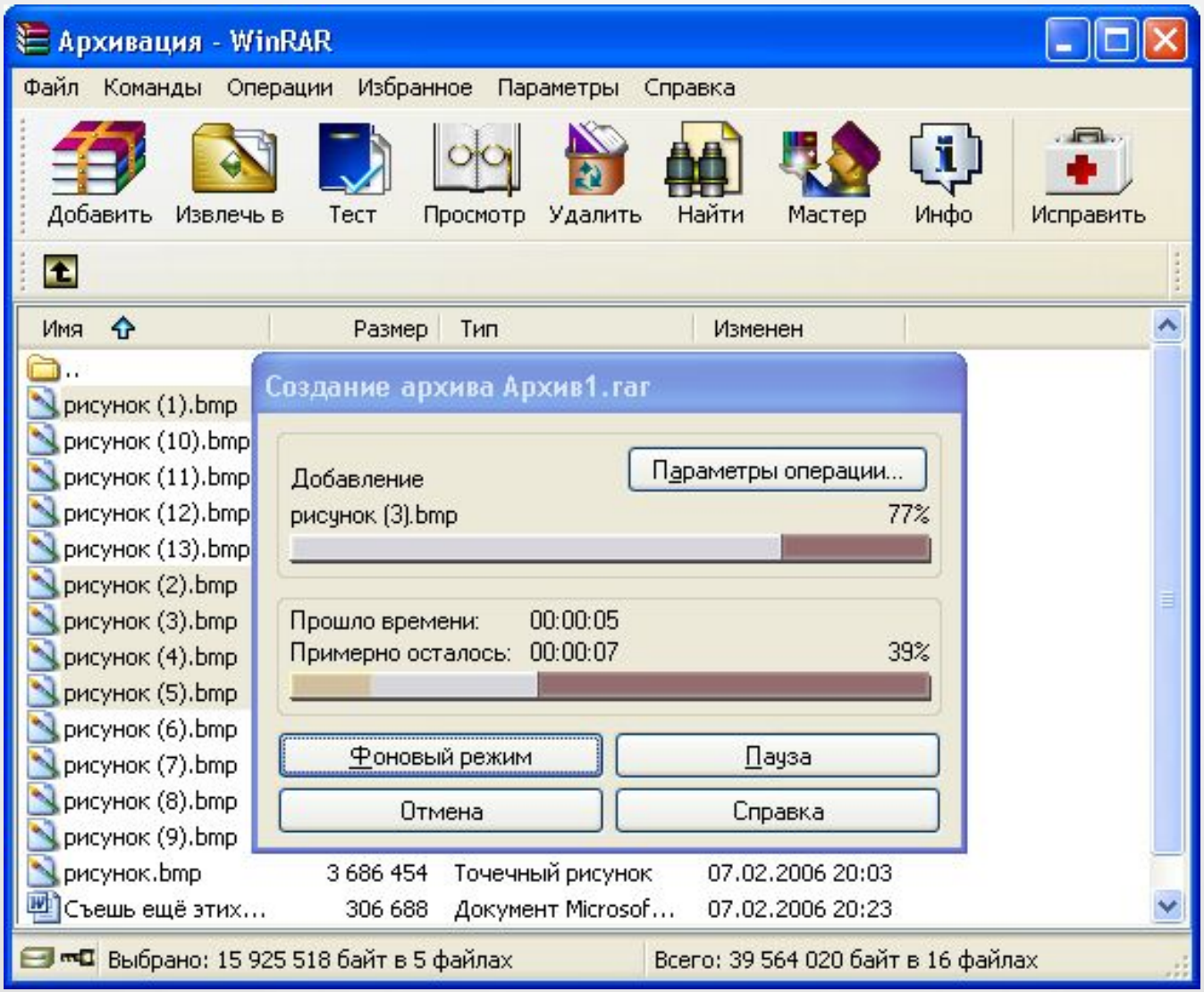
*

Архивация с помощью оболочки WinRAR

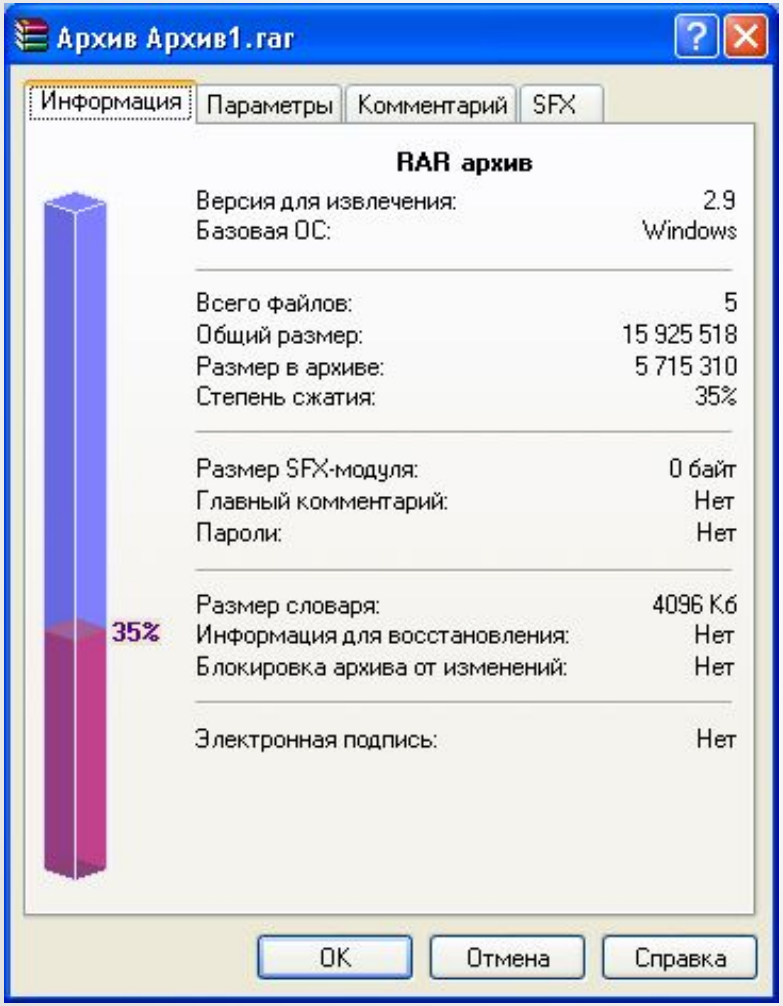


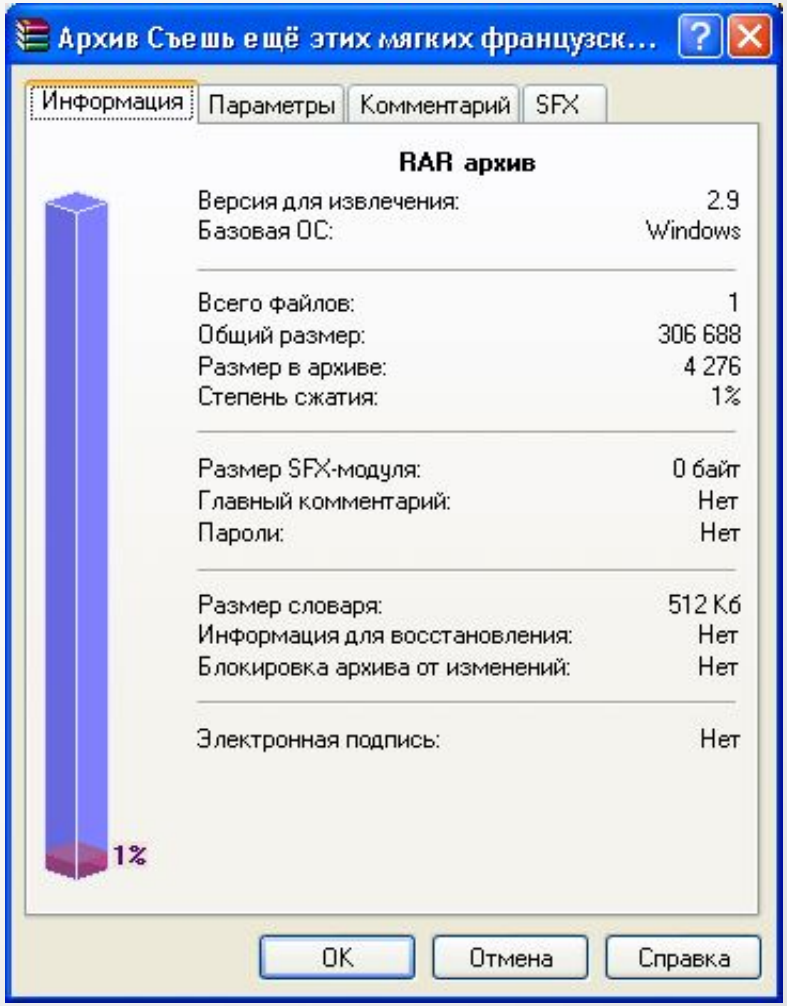
2

Архивация с помощью оболочки WinRAR



Информация об архиве

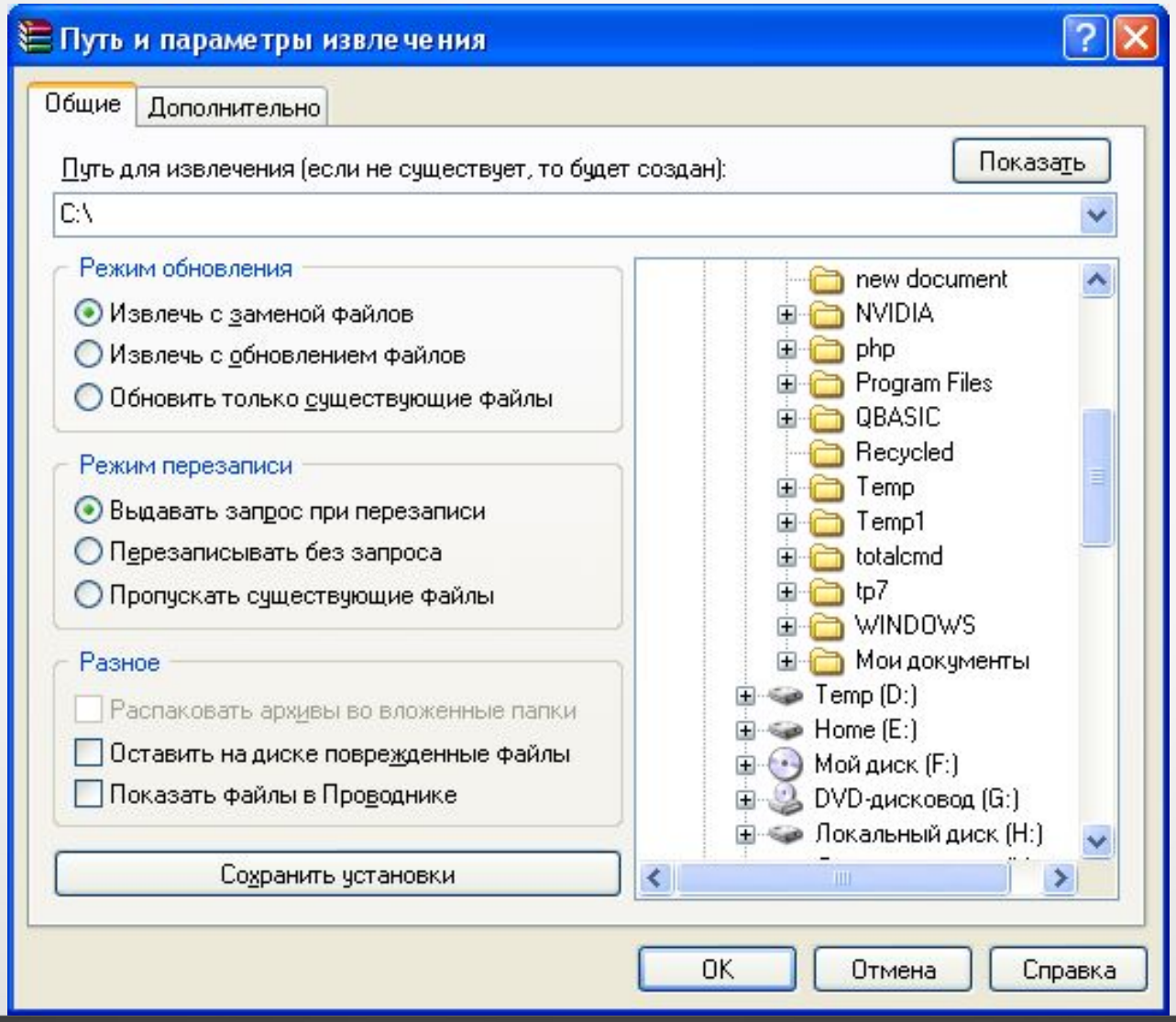




*

Распаковка файлов

Распаковка файлов



Вопросы:

Почему есть возможность уменьшать размер файлов?

Что такое архивация?

Какие файлы не имеет смысла архивировать?

Почему перед пересылкой текстового файла по электронной почте имеет смысл предварительно его упаковать в архив?

СРС

Составить таблицу сравнения свойств программ-архиваторов WinZip, WinRar, 7Zip, ARJ по следующему алгоритму:

1)Год создания

2)Алгоритм

3)Степень сжатия

4)Какие файлы сжимаются лучше при помощи этого архиватора

5)Интерфейс (графический, в виде командной строки)