Лекция №5

Файлы.
Типы файлов.
Структуры файлов.

Формат файла

Формат — спецификация структуры данных, записанных в компьютерном файле. Формат файла обычно указывается в его имени, как часть, отделённая точкой (обычно эту часть называют расширением имени файла, хотя, строго говоря, это неверно)

Определение формата файла

- Расширения
- Магические числа
- Метаданные
- MIME типы

Расширение

Некоторые операционные системы, например, CP/M, DOS, и Microsoft Windows используют для определения типа файла часть его имени, т.е. «расширение имени файла». В старых операционных системах это были три символа, отделённые от имени файла точкой (в файловых системах семейства FAT имя и расширение хранились отдельно, точка добавлялась уже на уровне ОС); в более новых системах расширение может являться просто частью имени, и тогда его длина ограничена только неиспользованной длиной имени (которая может составлять, например, 255 символов).

Несколько точек в имени

Магические числа

Широко используемый в UNIX-подобных операционных системах, заключается в том, чтобы сохранить в самом файле некое «магическое число» (magic number / сигнатуру) — последовательность символов, по которой может быть опознан формат файла. Первоначально этот термин использовался для специального набора 2-байтовых идентификаторов, сохраняемых в начале файла (эта практика перекочевала и в другие ОС, например, MZ в MS-DOS), однако, любая последовательность символов, характерная для данного формата, может быть использована как «магическое число».

Для определения формата файла служит команда file, которая использует файл /usr/share/misc/magic

Метаданные

Некоторые файловые системы позволяют сохранять дополнительные атрибуты для каждого файла, т. е. «метаданные». Эти метаданные можно использовать для хранения информации о типе файла. Такой подход используется в компьютерах Apple Macintosh. Метаданные поддерживаются такими современными файловыми системами как HPFS, NTFS, ext2, ext3 и другими. Недостатком этого метода является плохая переносимость — при копировании файлов между файловыми системами разных типов метаданные могут быть потеряны.

MIME

Типы данных, определённые стандартом MIME, широко используются в различных сетевых протоколах, однако в файловых системах они пока применяются редко.

Multipurpose Internet Mail Extensions (MIME (произн. «майм»), англ. Multipurpose Internet Mail Extension — многоцелевое расширение интернет-почты) — стандарт, описывающий передачу различных типов данных по электронной почте, а также, шире, спецификация для кодирования информации и форматирования сообщений таким образом, чтобы их можно было пересылать по Интернету.

Исполняемые файлы и код

Исполняемые
- EXE, COM
- BAT, CMD
- a.out, ELT

Исходный код
- ASM, C,CPP
- H, HPP (Заголовки / Headers)
- PAS, BAS
- JAVA, JS (JavaScript)
- PL (Perl), PHP, PY (Python)
- …

EXE

Portable Executable — (PE, произносится как [по́тэбл экзэкью́тэбл] — переносимый исполняемый) — формат исполняемых файлов, объектного кода и динамических библиотек, используемый в 32- и 64-битных версиях операционной системы Microsoft Windows. Формат PE представляет собой структуру данных, содержащую всю информацию, необходимую PE загрузчику для проецирования файла в память. PE представляет собой модифицированную версию COFF формата файла для Unix.

Расширения: .exe, .dll, .ocx, .sys, .scr, .drv, .cpl
Сигнатура: MZ(для совместимости), PE

COM

В системах DOS и в 8-битной CP/M COM-файл — простой тип исполняемого файла, размер которого не может превышать 64 Кбайт-256 байт.

Сигнатура: отсутствует

BAT / CMD

Batch file — Пакетный файл
Comand file — Командный файл

В Unix — обычно sh (shell). Позволяет в текстовом виде прописывать команды, которые должен выполнять командный процессор. Функциональность ограничена возможностями командного процессора и программ, которые можно запускать в командном режиме.

Unix shell — более развит. В Windows попытались сделать подобие в виде Windows PowerShell.

a.out / ELF

a.out

Расширения: нет, .o, .so

ELF (англ. Executable and Linkable Format — формат исполняемых и компонуемых файлов) — формат файлов, используемый во многих UNIX-подобных операционных системах, например, в GNU/Linux и Solaris, а также, после некоторой модификации ПО, — в некоторых мобильных телефонах компаний Siemens, Sony Ericsson, Motorola (платформа P2K) и во многих цифровых фотовидеокамерах (Olympus, Rekam и проч.).

Сигнатура: ELF (с первой позиции)

Документы

Текстовые документы
- TXT, DOC (прежний)
- DOC, RTF (Microsoft)
- PDF, PS, DjVu, FB2, EPUB (Книги)
- TEX, INFO
- DOCX, ODT / SXW

Презентации
- PPT, PPTX, ODP

Электронные таблицы
- XLS, XLSX, ODS
- CSV

TXT, DOC

Те́кстовый файл — компьютерный файл, содержащий текстовые данные, как правило, организованные в виде строк. Иногда конец текстового файла (особенно если в файловой системе не хранится информация о размере файла) также отмечается одним или более специальными знаками, известными как маркеры конца файла.

Текстовой файл может содержать как форматированный, так и неформатированный текст.

Различные операционные системы придерживаются своего представления перевода строки и конца файла. В UNIX перевод строки состоит из одного символа LF (0x0A), в Mac OS — из символа CR (0x0D), а в DOS и Windows перевод строки кодируется последовательностью двух символов: CR и LF. В DOS и Microsoft Windows конец файла кодируется символом 0x1A, а в UNIX символ конца файла не употребляется.

Помимо названных, в текстовых файлах применяются такие символы, как табуляция (0x09) и перевод страницы (0x0C).

DOC

Microsoft Word является наиболее популярным из используемых в данный момент текстовых процессоров, что сделало его бинарный формат документа стандартом де-факто, и многие конкурирующие программы имеют поддержку совместимости с данным форматом. Расширение «.doc» на платформе IBM PC стало синонимом двоичного формата Word 97—2000. Фильтры экспорта и импорта в данный формат присутствуют в большинстве текстовых процессоров. Формат документа разных версий Word меняется, различия бывают довольно тонкими. Форматирование, нормально выглядящее в последней версии, может не отображаться в старых версиях программы, однако есть ограниченная возможность сохранения документа с потерей части форматирования для открытия в старых версиях продукта.

Сигнатура: D0 CF 11 E0

RTF

Rich Text Format (RTF, «формат обогащённого текста» (rich с английского — богатый)) — проприетарный межплатформенный формат хранения размеченных текстовых документов, предложенный группами программистов, основавшими компании Microsoft и Adobe, как метатэговский формат для редактора Word в 1982 году. С тех пор спецификация формата несколько раз изменялась. RTF-документы поддерживаются всеми современными текстовыми процессорами. После разрыва отношений с Microsoft компания Adobe продолжила развитие метатэговского языка, заложенного в основу RTF, создав в 1985 году язык PostScript.
Хотя стандарт допускает 8-битное кодирование в отдельных случаях, текст в формате RTF обычно кодируется 7-битными символами. Это ограничило бы нас набором символов ASCII, но остальные символы можно кодировать с помощью escape-последовательностей.

Сигнатура: {\rtf

PDF

Portable Document Format (PDF) — кроссплатформенный формат электронных документов, созданный фирмой Adobe Systems с использованием ряда возможностей языка PostScript. В первую очередь предназначен для представления в электронном виде полиграфической продукции.
Формат PDF позволяет внедрять необходимые шрифты (построчный текст), векторные и растровые изображения, формы и мультимедиа-вставки. Поддерживает RGB, CMYK, Grayscale, Lab, Duotone, Bitmap, несколько типов сжатия растровой информации. Включает механизм электронных подписей для защиты и проверки подлинности документов.

Сигнатура: %PDF

PS

PostScript (Постскрипт) — язык описания страниц, в основном используемый в настольных издательских системах.

PostScript — больше, чем типичный язык управления принтером, он является полнофункциональным языком программирования. Многие прикладные программы могут преобразовать документ в PostScript-программу, при выполнении которой будет получен начальный документ. Эта программа может быть послана непосредственно на принтер с поддержкой PostScript или преобразована интерпретатором PostScript в другой формат (для принтеров без поддержки PostScript), или результат её выполнения интерпретатором может быть показан на экране. Так как исходная PostScript-программа одна и та же, PostScript называется независимым от устройства.

Сигнатура: %!PS

DjVu

DjVu (от фр. déjà vu — «уже виденное») — технология сжатия изображений с потерями, разработанная специально для хранения сканированных документов — книг, журналов, рукописей и прочее, где обилие формул, схем, рисунков и рукописных символов делает чрезвычайно трудоёмким их полноценное распознавание. Также является эффективным решением, если необходимо передать все нюансы оформления, например, исторических документов, где важное значение имеет не только содержание, но и цвет и фактура бумаги; дефекты пергамента: трещинки, следы от складывания; исправления, кляксы, отпечатки пальцев и т. д.
Формат оптимизирован для передачи по сети таким образом, что страницу можно просматривать ещё до завершения загрузки файла. DjVu-файл может содержать текстовый (OCR) слой, что позволяет осуществлять полнотекстовый поиск по файлу.
Расширения: .djvu, .djv

XML

XML (англ. eXtensible Markup Language — расширяемый язык разметки; произносится [экс-эм-э́л]) — свод общих синтаксических правил. XML — текстовый формат, предназначенный для хранения структурированных данных, для обмена информацией между программами, а также для создания на его основе более специализированных языков разметки (например, XHTML). XML является упрощённым подмножеством языка SGML.

Расширение: .xml

Сигнатура: <?xml

FB2

FictionBook — формат представления электронных версий книг в виде XML-документов, где каждый элемент книги описывается своими тегами. Стандарт призван обеспечить совместимость с любыми устройствами и форматами.

Правильно подготовленный электронный текст в формате FictionBook содержит в себе всю необходимую информацию о книге: структурированный текст, иллюстрации, информацию об авторе и издании, но не содержит информацию о внешнем виде документа. Как будет выглядеть текст, полученный из формата .fb2, зависит либо от настроек программы-просмотрщика этого формата, либо от параметров, заданных при конвертации файла в другой формат.

Расширение: fb2, fb2.zip

EPUB

Electronic Publication (ePub) — открытый формат электронных версий книг, разработанный Международным форумом по цифровым публикациям IDPF. Файлы в этом формате имеют расширение .epub. Формат позволяет издателям производить и распространять цифровую публикацию в одном файле, обеспечивая совместимость между программным и аппаратным обеспечением, необходимым для воспроизведения цифровых книг и других публикаций с плавающей вёрсткой.
Zip-архив контейнера ePub содержит тексты в форматах xHTML, HTML или PDF, описание издания в XML, рядом в папках — графика, включая векторную (SVG), и встроенные шрифты, таблицы стилей и т. д.

Расширение: epub

TeX

TEX (обычным текстом — TeX; произносится «тех») — система компьютерной вёрстки, разработанная американским профессором информатики Дональдом Кнутом в целях создания компьютерной типографии. В неё входят средства для секционирования документов, для работы с перекрёстными ссылками. Многие считают TeX лучшим способом для набора сложных математических формул.

Расширение: .tex

TexInfo

Texinfo (рус. Текинфо) — свободная система документирования и язык разметки, позволяющие создавать документы в разных форматах из одного исходного текста.
Исходные файлы Texinfo-документов представляют собой простой текст, размеченный при помощи специальных команд, начинающихся со знака @ (например, @contents или @titlepage). Файлы Texinfo обычно имеют расширение .texi, реже .txi.

При помощи утилит makeinfo, texi2dvi и texi2pdf, входящих в Texinfo, из исходных файлов генерируется документация в других форматах.

ODT

OpenDocument Format, ODF (от англ. OASIS Open Document Format for Office Application — рус. открытый формат документов для офисных приложений) — открытый формат файлов документов для хранения и обмена редактируемыми офисными документами, в том числе текстовыми документами (такими как заметки, отчёты и книги), электронными таблицами, рисунками, базами данных, презентациями.
Стандарт был разработан индустриальным сообществом OASIS и основан на XML-формате. 1 мая 2006 года принят как международный стандарт ISO/IEC 26300.
Microsoft Office 2007 поддерживает формат OpenDocument, начиная с SP2.

DOCX

Office Open XML (OOXML, DOCX, проект ISO/IEC IS 29500:2008) — серия форматов файлов для хранения электронных документов пакетов офисных приложений — в частности, Microsoft Office. Формат представляет собой zip-архив, содержащий текст в виде XML, графику и другие данные, которые могут быть переведены в последовательность битов (сериализованы) с применением защищённых патентами двоичных форматов.

Две разные версии OOXML определены в ECMA-376 и в ISO 29500:2008. Полная поддержка формата ISO 29500 ожидалась в Microsoft Office 2010.

В марте 2008 года спецификация была принята как будущий стандарт ISO/IEC 29500.

Презентации

PPT / PPTX / ODP

Электронные таблицы

XLS / XLSX / ODS

CSV

CSV (от англ. **Comma-Separated Values** — значения, разделённые запятыми) — текстовый формат, предназначенный для представления табличных данных. Каждая строка файла — это одна строка таблицы. Значения отдельных колонок разделяются разделительным символом (*delimiter*) — запятой (,). Однако, большинство программ вольно трактует стандарт CSV и допускают использование иных символы в качестве разделителя. В частности в локалях, где десятичным разделителем является запятая, в качестве табличного разделителя, как правило, используется точка с запятой. Значения, содержащие зарезервированные символы (пробел, запятая, точка с запятой, новая строка) обрамляются двойными кавычками ("); если в значении встречаются кавычки — они представляются в файле в виде двух кавычек подряд.

CSV

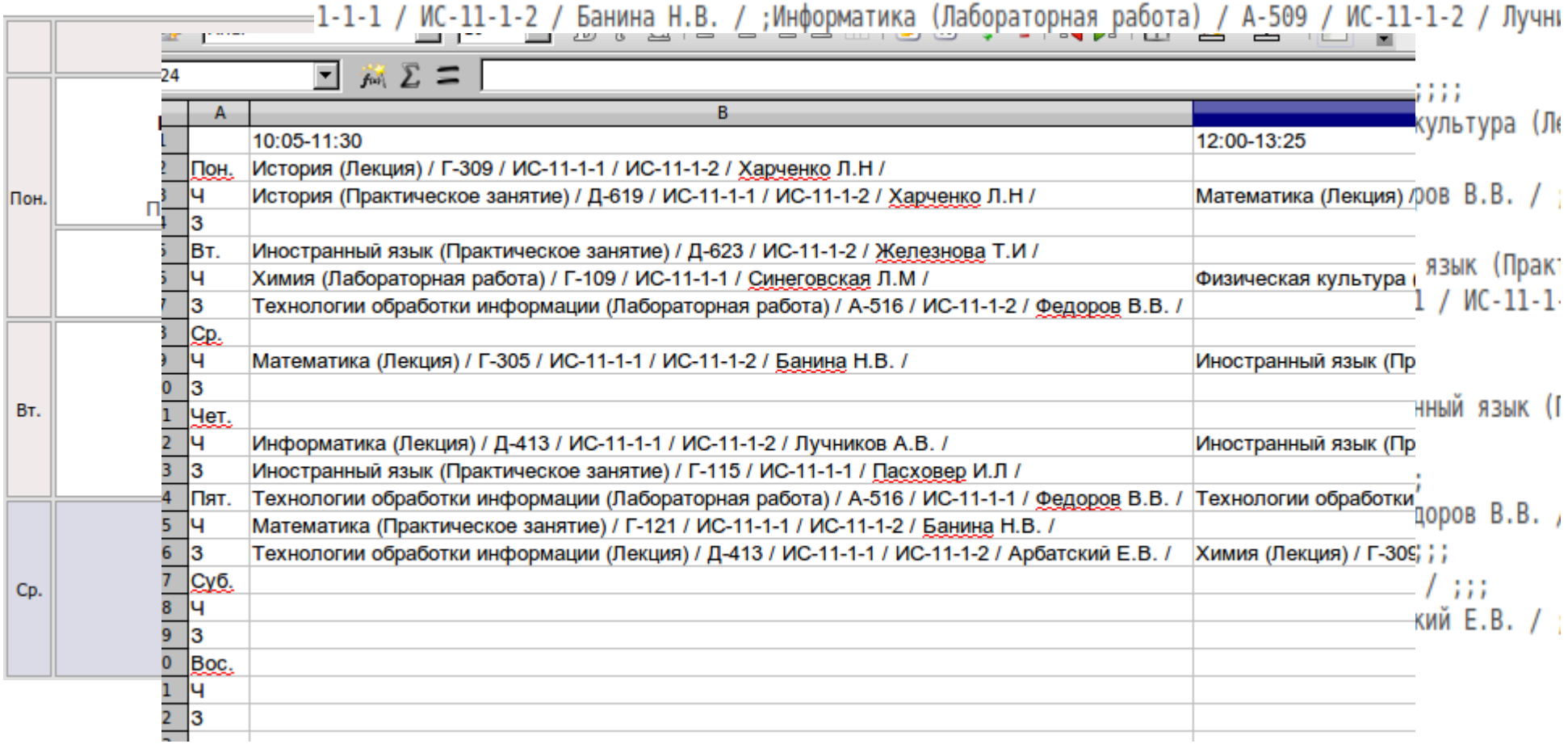
```
1965;Пиксел;E240 — формальдегид (опасный консервант)!;"красный, зелёный, битый";3000,00
1965;Мышка;"А правильней ""Использовать Ёлочки""";;4900,00
"Н/д";Кнопка;Сочетания клавиш;"MUST USE! Ctrl, Alt, Shift";4799,00
```

Результирующая таблица:

1965	Пиксел	E240 — формальдегид (опасный консервант)!	красный, зелёный, битый	3000
1965	Мышка	А правильней "Использовать Ёлочки"		4900
Н/д	Кнопка	Сочетания клавиш	MUST USE! Ctrl, Alt, Shift	4799

CSV

```
eugine@eugene-945P-S3:~/Downloads$ iconv -f cp1251 IS-11-1.csv
;10:05-11:30;12:00-13:25;13:35-15:00;15:10-16:35;16:45-18:10;18:20-19:45;19:55-21:20
Пон.;;История (Лекция) / Г-309 / ИС-11-1-1 / ИС-11-1-2 / Харченко Л.Н / ;;;;;
Ч;;;История (Практическое занятие) / Д-619 / ИС-11-1-1 / ИС-11-1-2 / Харченко Л.Н / ;Математика
1-1-1 / ИС-11-1-2 / Банина Н.В. / ;Информатика (Лабораторная работа) / А-509 / ИС-11-1-2 / Лучни
```



	A	B	
1		10:05-11:30	12:00-13:25
2	Пон.	История (Лекция) / Г-309 / ИС-11-1-1 / ИС-11-1-2 / Харченко Л.Н /	
3	Ч	История (Практическое занятие) / Д-619 / ИС-11-1-1 / ИС-11-1-2 / Харченко Л.Н /	Математика (Лекция) /
4	З		
5	Вт.	Иностранный язык (Практическое занятие) / Д-623 / ИС-11-1-2 / Железнова Т.И /	
6	Ч	Химия (Лабораторная работа) / Г-109 / ИС-11-1-1 / Синеговская Л.М /	Физическая культура (
7	З	Технологии обработки информации (Лабораторная работа) / А-516 / ИС-11-1-2 / Федоров В.В. /	
8	Ср.		
9	Ч	Математика (Лекция) / Г-305 / ИС-11-1-1 / ИС-11-1-2 / Банина Н.В. /	Иностранный язык (Пр
10	З		
11	Чет.		
12	Ч	Информатика (Лекция) / Д-413 / ИС-11-1-1 / ИС-11-1-2 / Лучников А.В. /	Иностранный язык (Пр
13	З	Иностранный язык (Практическое занятие) / Г-115 / ИС-11-1-1 / Пасховер И.Л /	
14	Пят.	Технологии обработки информации (Лабораторная работа) / А-516 / ИС-11-1-1 / Федоров В.В. /	Технологии обработки
15	Ч	Математика (Практическое занятие) / Г-121 / ИС-11-1-1 / ИС-11-1-2 / Банина Н.В. /	
16	З	Технологии обработки информации (Лекция) / Д-413 / ИС-11-1-1 / ИС-11-1-2 / Арбатский Е.В. /	Химия (Лекция) / Г-309
17	Суб.		
18	Ч		
19	З		
20	Вос.		
21	Ч		
22	З		

Архиваторы

- ZIP
- RAR
- 7Z
- TAR, GZIP, BZ2
- HA, LHA, …

ZIP

ZIP — популярный формат сжатия данных и архивации файлов. Файл в этом формате обычно имеет расширение .zip

Формат ZIP был первоначально создан Филом Кацем, основателем компании PKWARE, в ответ на правовое преследование компанией Software Enhancement Associates (SEA), защищавшей своё изобретение — формат архивирования ARC.

Сигнатура: PK

RAR

RAR — распространённый проприетарный формат сжатия данных и программа-архиватор.

Формат разработан российским программистом Евгением Рошалом (отсюда и название RAR: Roshal Archiver). Он написал программу-архиватор для упаковки/распаковки RAR, изначально под DOS, затем и для других операционных систем. Версия для Microsoft Windows распространяется в составе многоформатного архиватора с графическим интерфейсом WinRAR.

Программа распространяется как условно-бесплатное программное обеспечение (shareware).

Расширения: .rar, .r00 …
Сигнатура: Rar

7Zip

7-Zip — свободный файловый архиватор с высокой степенью сжатия данных. Поддерживает несколько алгоритмов сжатия и множество форматов данных, включая собственный формат 7z с высокоэффективным алгоритмом сжатия LZMA.

Поддерживаемые форматы:

упаковка и распаковка: 7z, BZIP2 (BZ2, TB2, TBZ, TBZ2), GZIP (GZ, TGZ), TAR, ZIP, XZ;

только распаковка: 001, ACE, ARJ, CAB, CHM, CPIO, DEB, DMG, FLV, ISO, JAR, LHA, LZH, LZMA, LZO (TZO), MSI, NSIS, PE, RAR, RPM, SWF, SWM, VHD, WIM, XAR, Z (TAZ); FAT, HFS, MBR, NTFS, UDF, SquashFS, CramFS

ARJ

ARJ — файловый архиватор. Разработан Робертом К. Джангом (Robert K. Jung). (Происхождение наименования ARJ: Archiver Robert Jung). ARJ версии 1.00 был выпущен в феврале 1991 г. под лицензией shareware.

ARJ компрессия подобна PKZIP 1.02

Существует также версия ARJ с открытым исходным кодом, доступная под более, чем десятью операционными системами.

Сигнатура: `

TAR

tar (англ. tape archive) — формат битового потока или файла архива, а также название традиционной для Unix программы для работы с такими архивами. Программа tar была стандартизирована в POSIX.1-1998, а также позднее в POSIX.1-2001. Первоначально программа tar использовалась для создания архивов на магнитной ленте, а в настоящее время tar используется для хранения нескольких файлов внутри одного файла, для распространения программного обеспечения, а также по прямому назначению — для создания архива файловой системы. Одним из преимуществ формата tar при создании архивов является то, что в архив записывается информация о структуре каталогов, о владельце и группе отдельных файлов, а также временны́е метки файлов.

Сигнатура: u s t a r \0 0 0 at byte 257

GZIP

gzip (сокращение от GNU Zip) — утилита сжатия и восстановления (декомпрессии) файлов, использующая алгоритм DEFLATE. Используется в основном в UNIX-системах, в ряде которых является стандартом де-факто для сжатия данных.

В соответствии с традициями UNIX-программирования, gzip выполняет только две функции: сжатие и распаковка одного файла, он не умеет упаковывать несколько файлов в один архив.

С другой стороны, указанная особенность даёт gzip возможность работать с непрерывным потоком данных, упаковывая/распаковывая их «на лету».

BZIP2

bzip2 — бесплатная свободная утилита командной строки с открытым исходным кодом для сжатия данных, реализация алгоритма Барроуза — Уилера.
В соответствии с традициями UNIX-программирования, bzip2 единовременно выполняет только одну функцию: сжатие или распаковку одного файла.
bzip2 сжимает большинство файлов эффективнее, но медленнее, чем более традиционные gzip или ZIP.

Расширение: .bz2

Сигнатура: BZh

Медиа файлы

Графика

- BMP, GIF, PNG
- JPEG, TIFF
- RAW

3D

- DWG (AutoCAD)
- MA (Maya)
- X3D, 3DS, BLEND, …

Аудио

- WAV
- MP3, AAC, APE, FLAC
- OGG, WMA, …

Видео

- AVI, 3GP, FLV
- MPEG, MKV, QuickTime
- MP4, WMV