CEP - UP - UTFPR - PUCPr - UFPR –
17/06/2024 - 17:11:23.4
Prof Dr P Kantek (pkantek@gmail.com)
Objetos binários grandes vivxq10a, V: 1.01 76506
CARLOS EDUARDO ANTAO 24CC2301 - 1
entregar ate 4/julho _____ / _____ / _____

Objetos binários grandes - BLOBs

O objetivo desta atividade e conhecer e manusear os chamados objetos binários grandes: imagens, vídeos, documentos, músicas, etc. Tais objetos quando estão na memória do computador são conhecidos como BLOBs e quando vão para uma memória não volátil (pendrives, discos, SSDs, nuvem, ...) passam a ser arquivos.

Resultados da digitalização crescente da nossa sociedade, precisam ser conhecidos para deles se extraírem todos os benefícios que eles apresentam.

Eis alguns raciocínios: a documentação em papel de um avião boeing 747 ocuparia 5 aviões 747 para ser transportada. Alguém ainda compra enciclopédias em papel ? Lembram de quando as fotografias eram analógicas ? Um filme Ektachrome (36 fotos) custava bem caro. Vou chutar uns 150 reais, o que daria quase 5 reais/foto, só o filme, sem contar a revelação e as ampliações. Quando o I.R. era em papel, a receita precisava montar uma operação de guerra, envolvendo todos os bancos, apenas para receber as declarações, e isso que era mais ou menos 1/3 do número de contribuintes atual. E as músicas: durante muitos anos, no dia do meu pagamento eu ia a uma loja de discos (?) e comprava 1 único LP, ($\pm$ 30 minutos de música) era o que o meu dinheiro alcançava. Aliás, falando de música, vale comentar a ascensão e queda do formato CD. Ele passou de maravilha tecnológica em meados dos 80 a algo obsoleto em meados dos 2010. Do céu ao inferno em 30 anos. E, se perder em uma cidade estranha, tendo apenas o guia 4 Rodas para ajudar ? Uma sensação indescritível.

Vamos estudar mais de perto alguns BLOBs, lembrando que qualquer um pode criar e chamar de seu, um BLOB. Pela sua definição (grande bloco de dados binários com uma lógica interna de organização e acesso) vê-se que isto é perfeitamente possível. Assim, vamos nos limitar a BLOBs padrão como formatos e conceitos conhecidos e publicados.

Imagem É difícil contar esta história. Ela começa no daguerrótipo de meados do século 19, através da fixação de imagens em placas de vidro. Segue pelas imagens de TV que eram capturadas e transmitidas eletronicamente. Seu registro foi feito em gravações magnéticas analógicas. Uma nova necessidade surgiu na obtenção de fotos na Lua e além, quando não havia como retornar filmes. A propósito, um pouco antes disto, a vigilância eletrônica na Guerra Fria era muito custosa: satélites fotografavam o inimigo e precisavam ejetar os filmes (que acabavam rápido) ao sobrevoar território amigo. A primeira foto da lua, transmitida eletronicamente foi feita pela espaçonave Ranger em 1964, 17 minutos antes dela se espatifar na superfície lunar. Quando a Internet começou (por volta de 1990) a necessidade de manusear imagens sofreu uma mudança: antes o gerador e o visualizador eram de mesma origem, pelo que não havia necessidade de padronização. Com a Internet essa história mudou. Uma das primeiras respostas a esta demanda foi o formato BMP (Bitmap Image File). Embora originalmente proposto pela Microsoft e IBM, teve seu padrão publicado e virou de fato, um formato livre. Sua origem é a bio-física dos olhos humanos (3 canais separados para Red, Green e Blue), além de um truque bem legal chamado mapeamento pelo qual se negocia tamanho do arquivo VERSUS menor quantidade de cores. A principal inovação do BMP foi criar, publicar e obedecer a um padrão. Parece pouco, mas não é. A chave da popularização e do sucesso sempre é um PADRãO. Anote isso na sua cabeça e nunca mais esqueça. O padrão pode ser ruim ou bom (neste caso não é muito), mas é um padrão. A imagem é uma matriz numérica (outra inovação entusiasmadora), logo sujeita a sofrer a ação maravilhosa de qualquer matemática.

Padrão BMP O arquivo começa com um bloco de controle de 54 bytes, que contém o identificador 'BM', o tamanho do objeto (arquivo), onde a imagem começa, L=largura e A=altura da imagem, profundidade). Depois, dependendo da profundidade, pode haver uma tabela de cores. Finalmente a imagem na forma de $L \times A$ se for uma imagem mapeada ou 3 tabelas de $L \times A$, uma para cada cor se for uma imagem true-color. Veja na imagem a seguir estes conceitos

```
Exemplo:
       0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
      -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
0000  42 4D 62 05 00 00 00 00 00 00 36 04 00 00 28 00 BM.....6...(
0010  00 00 12 00 00 00 0F 00 00 00 01 00 08 00 00 00 ................
0020  00 00 60 01 00 00 00 00 00 00 00 00 00 00 00 01 ..`.............
0030  00 00 00 00 00 00 00 00 FF FF 00 80 80 00 FF FF ................
0040  FF 00 FF 00 FF 00 00 00 00 FF 00 00 00 00 00 FF ................
0050  00 00 FF FF 00 00 FF 00 00 00 09 09 09 00 0A 0A ................
0060  0A 00 0B 0B 0B 00 0C 0C 0C 00 0D 0D 0D 00 0E 0E ................
0070  0E 00 0F 0F 0F 00 10 10 10 00 11 11 11 00 12 12 ................
...
0410  F6 00 F7 F7 F7 00 F8 F8 F8 00 F9 F9 F9 00 FA FA ................
0420  FA 00 FB FB FB 00 FC FC FC 00 FD FD FD 00 FE FE ................
0430  FE 00 FF FF FF 00 01 01 03 01 01 01 08 01 01 01 ................
0440  01 01 02 01 01 06 01 01 00 00 04 04 03 04 04 04 ................
0450  04 04 04 04 04 04 04 04 04 04 04 04 00 00 01 01 ................
```

Os problemas do BMP incluem: nenhuma compressão, o que graças ao fenômeno da *localidade de referência espacial* é uma tragédia. Também não prevê nenhum tipo de animação, o que faz falta na Internet. Não sabia (hoje sabe) lidar com transparência.

Um concorrente importante do BMP (lá nos primórdios) foi o formato GIF que admite compressão (usando o belo algoritmo LZW) e também admite uma forma rudimentar de animação. O problema do GIF é que ele foi e é um formato proprietário. Isto deu origem ao formato PNG que é praticamente o mesmo, mas livre.

A compressão poderia ser muito melhor, se se admitisse algum nivel de degradação na reconstituição da imagem após sua transformação em BLOB. Isto foi conseguido no padrão JPEG, que ao admitir degradações (sob controle), gera compressões muito maiores do que GIF ou similares. Outra vantagem é a possibilidade de criação em etapas das imagens (coisa que o BMP nunca ofereceu). O padrão foi formalizado pela ISO em 1991, mas na época disparou uma briga por patentes e direitos. Nos EUA, as últimas patentes expiraram em 2004. A compressão aqui é pelo uso da Transformada Discreta do Coseno. A sua operação está além do nosso interesse (querendo, olhe VIVX690c ou 690p), mas em resumo: a imagem é quebrada em blocos de 8×8; esta matriz é multiplicada pela matriz DCT, o que passa do domínio do pixel para o domínio da frequência (Transformada de Fourier); os coeficientes são quantizados; os coeficientes são comprimidos. Depois, se faz o inverso e *voilà*.

Som O MP3, sigla para MPEG-1 Audio Layer III, é um formato de arquivo de áudio digital que revolucionou a forma como consumimos e compartilhamos música.

Desenvolvido no início da década de 1990 pelo Moving Picture Experts Group (MPEG), o MP3 se tornou o formato dominante para música digital, utilizado em players de música portáteis, computadores, smartphones e serviços de streaming online. O MP3 utiliza uma técnica de compressão com perda chamada percepção auditiva psicoacústica para reduzir o tamanho dos arquivos de áudio. Essa técnica remove informações da faixa de áudio que o ouvido humano provavelmente não consegue perceber, resultando em arquivos menores sem comprometer significativamente a qualidade da música para a maioria dos ouvintes. Apesar da compressão, o MP3 oferece uma qualidade de som bastante boa, especialmente em taxas de bits mais altas (como 128 kbps ou superior). Tamanho de Arquivo Reduzido: Os arquivos MP3 são significativamente menores do que os arquivos de áudio não compactados, como CDs, permitindo fácil armazenamento e compartilhamento de música digital.

O MP3 teve um impacto profundo na indústria da música e na forma como consumimos música gerando um declínio das vendas de CDs: Os consumidores passaram a baixar e compartilhar músicas digitalmente. O MP3 impulsionou o crescimento da música digital, com o surgimento de serviços de streaming online como Spotify, Apple Music e Deezer. O MP3 abriu caminho para novos modelos de negócios na indústria da música, como a venda de músicas online e assinaturas de serviços de streaming.

Apesar de seu sucesso, o MP3 também enfrentou alguns desafios: Em taxas de bits mais baixas (como 64 kbps), a qualidade de som do MP3 pode ser perceptívelmente inferior à do áudio não compactado. Propriedades Intelectuais: Questões de direitos autorais e pirataria surgiram com a proliferação de arquivos MP3 compartilhados online. Formatos Alternativos: Novos formatos de áudio digital, como FLAC e AAC, oferecem maior qualidade de som ou taxas de compressão mais eficientes, mas ainda não alcançaram o mesmo nível de popularidade do MP3.

Outros padrões: MID (notação musical sintética), WAV (nenhuma perda de qualidade, logo arquivos grandes), além do formato OGG, da plataforma APPLE.

Vídeo No vídeo, estamos diante de necessidade maior de compressão. Se uma imagem estática já é grande, imagine precisar de 30 imagens estáticas por segundo. Agora, a localidade de referência ganha um novo viés, a l.r. no tempo. Ou seja, se imaginarmos um vídeo como um array 3-d (dimensões tempo, altura e largura), podemos comprimir em 2 dessas dimensões.

Aqui a história começa com os formatos analógicos (VHS, Betamax), segue pelo CD-ROM e VCD, e chega ao DVD. Depois vem os formatos AVC (padrão H.264), depois HEVC (H.265), VP9 (formato livre e aberto da Google) e AV1 (formato livre e aberto, melhor que o H.265). Diferentemente da imagem (e semelhantemente ao som), aqui existe muita preocupação com a latência (atraso no envio dos pacotes) que impacta diretamente a sincronia. O padrão H.264 utiliza timestamps e buffers para sincronizar áudio e vídeo. Em caso de perda de pacotes, mecanismos de recuperação de erros como o Concealment Motion Vector (CMV) podem ser utilizados para minimizar o impacto na sincronia. O CMV divide o vídeo em macroblocos. Quando há perda de um pacote contendo os vetores de movimento de um determinado macrobloco, o CMV busca os macroblocos vizinhos ($\uparrow, \downarrow, \leftarrow \rightarrow$) e ve se algum deles tem movimento similar ao esperado do macrobloco perdido. Daí ele "empresta" vetores e reconstitui o movimento ausente.

PDF O Portable Document Format (PDF), criado pela Adobe em 1993, se tornou um dos formatos de arquivo mais utilizados no mundo, presente em diversos setores e aplicações. Mais detalhes: vivx680a. Um documento PDF pode ser aberto e visualizado em qualquer dispositivo com um leitor de PDF instalado, independentemente do sistema operacional ou hardware utilizado. Isso garante que a formatação original do documento é sempre a mesma, independentemente da plataforma em que ele é aberto. Isso é crucial para documentos que exigem alta fidelidade visual, como apresentações, relatórios e publicações. O PDF oferece recursos de segurança robustos, como criptografia e assinaturas digitais, que protegem o conteúdo contra acessos não autorizados, modificações e falsificações. Isso torna o PDF ideal para documentos confidenciais e contratos legais. Até porque ele pode ser acompanhado de uma assinatura digital (*hash*). Suporta recursos de acessibilidade, como tags de estrutura e legendas para imagens, que facilitam o acesso ao conteúdo do documento por pessoas com deficiência visual ou outras necessidades especiais. Ao contrário de outros formatos de arquivo que podem ser corrompidos ou se tornar obsoletos com o tempo, o PDF é um formato durável e estável. Pense na economia de árvores: imagine se todos os PDFs existentes fossem para o papel. Ele é ideal para compartilhar documentos com outras pessoas, pois garante que a formatação original seja preservada e o conteúdo seja visualizado corretamente em diferentes dispositivos. O PDF é uma ótima opção para armazenar documentos importantes de forma segura e durável, pois protege o conteúdo contra modificações e garante que ele possa ser acessado no futuro. Revistas, livros, relatórios e outros materiais de publicação podem ser distribuídos no formato PDF, garantindo uma experiência de leitura consistente em diferentes dispositivos. **Outros:** Deixo de tratar aqui, por absoluta falta de espaço, outros BLOBs como exames médicos, arquivos de telescópios (vivx760), ...

Olhando dentro de um BLOB

Agora e hora de examinar com um "microscópio" as partes internas de um desses arquivos. O escolhido pela importância é o arquivo PDF. Pode parecer pouco, mas esta ideia simples, já salvou centenas de milhares (milhões ?) de árvores. Já são muitos os sistemas de informação que não geram mais saídas em papel, limitando-se a criar arquivos PDF.

Um arquivo PDF descreve como uma página ficará depois de impressa e ele é melhor do que imagem dessa página pelas seguintes razões

- ⋆ A imagem como tal sempre é muito maior do que sua descrição. A razão para isso é a evidente redundância de qualquer imagem.

⋆ A imagem como tal é fixa e só pode ser rotacionada, redimensionada ou de alguma maneira modificada à custa de processamento e perda de qualidade.

⋆ Sobre uma imagem não há como manipular o conteúdo, por exemplo conduzindo uma busca textual (^F alguma coisa ou então ^C e ^V).

⋆ Alterar uma única palavra sobre uma imagem exige habilidades de pintura. Sobre a descrição de uma página é tarefa simples: basta trocar uma palavra pela outra.

⋆ Imagine uma imagem enviada para um monitor (72 DPP) ou a mesma imagem enviada para uma impressora (300 DPP). Não podem ser a mesma imagem.

A versão 1.0 do padrão PDF nasceu em 1993 quando a Adobe lançou 2 programas: o Distiller (para gerar PDF) e o Reader (para ler): ambos pagos. A coisa começou a mudar quando o departamento de rendas americano comprou uma licença que permitia aos contribuintes baixar o Reader. Rapidamente a Adobe deu-se conta de que seria melhor para ela distribuir gratuitamente o Reader. Nos dez anos seguintes o PDF acabou se tornando o padrão de fato de descrição de imagens.

Apresenta as seguintes vantagens:

⋆ acesso aleatório: qualquer página pode ser montada independente das demais.

⋆ linearização: todos os elementos necessários para montar uma página estão juntos.

⋆ atualização incremental: mudanças ficam registradas no fim do arquivo. Vantagem: opção de desfazer ilimitada e rápida atualização. Desvantagem: o arquivo não pára de crescer.

⋆ fontes incorporadas: o destino sempre será montado corretamente. A fonte é desbastada de tudo que não é necessário.

⋆ padronização ISO: em 2008 a ISO abraçou o padrão Adobe PDF 1.7, o que o transformou em padrão aberto.

⋆ compatibilidade para trás e para a frente: Qualquer leitor lê as versões antigas e deve ler as futuras, já que todos os leitores ignoram o que não conhecem.

⋆ integrado a todos os principais browsers do protocolo `HTTP`.

Um arquivo PDF

Cabeçalho Formado por 2 linhas: a primeira identifica o arquivo como um PDF e informa a versão PDF em que ele foi gerado. A segunda linha inclui caracteres que não podem ser impressos (para avisar eventuais leitores deste fato). Exemplo:

```
%PDF-1.0
%âã
```

Corpo Um conjunto de nodos de um grafo, contendo as páginas, conteúdo gráfico, textos, sempre na forma de objetos. Cada objeto começa com um número de objeto (iniciando em 1), um número de geração (sempre 0) e a palavra chave `obj`. O objeto termina pela palavra chave `endobj`. Exemplo:

```
1 0 obj     -- descrevendo o objeto 1
<<          -- começa um dicionário
 /Kids [2 0 R]    -- /Kids tem o valor 2 0 R
 /Count 1         -- só um
 /Type /Pages     -- do tipo : página
>>               -- fim do dicionário
endobj            -- fim do objeto
```

Tabela de Referência Cruzada Lista o deslocamento em bytes de cada objeto no corpo do arquivo. Exemplo

```
0 6          -- seis entradas na tabela
0000000000  65535  f   -- entrada especial
0000000015  00000  n   -- obj1 no desloc 15
0000000074  00000  n   -- obj2 no desloc 74
...
```

Rodapé Informações e metadados para aumentar a performance de acesso. A primeira entrada é `trailer`. Depois o dicionário de trailer que deve apresentar ao menos a entrada `\Size` dizendo quantos itens há na tabela de referência cruzada e `\Root` dizendo qual objeto é o catálogo do documento. Depois vem uma linha que só contém `startxref` seguida por uma única linha indicando o deslocamento em bytes utilizado no inicio da tabela de referência cruzada do arquivo. Termina tudo a linha `%%EOF`. Exemplo:

```
trailer   -- palavra chave
<<   -- começa um dicionário
/Root 5 0 R   -- o catálogo é o objeto 5
/Size 6       -- tem o tamanho 6
>>   -- fim do dicionário
startxref
459  -- deslocamento da tabela de ref cruzada
%%EOF   -- fim de arquivo
```

Componentes

Uma descrição PDF suporta 5 componentes básicos: i. Números inteiros e reais (mas não exponenciais); ii. strings que são escritos entre parênteses; iii.Nomes, que sempre começam por uma barra normal; iv.Os valores booleanos `true` e `false` e v. O componente nulo (`null`). Aceita também 3 componentes compostos: i. Arrays, que são uma coleção ordenada de outros componentes, escritos entre colchetes; ii.Dicionários que são uma lista não ordenada de duplas: nome e componente. Começam por `<<` e terminam por `>>`. Finalmente iii. Fluxos, que armazenam dados binários, sempre acompanhados de um dicionário que descreve seus atributos, como comprimento e parâmetros de compactação, por exemplo.

☞ Para você fazer

As coisas em um arquivo PDF são separadas por um divisor de linhas. Pode ser o caracter $X'10'$, ou o $X'13'$ ou uma combinação de ambos. Ao examinar um PDF em hexadecimal, acostume-se a separar as linhas marcando este caractere. Isto posto, os componentes principais do arquivo PDF são:

- Header: identificação PDF e uma coleção de letras acentuadas
- Corpo: objetos numerados a partir de 1 sequencialmente, e identificados pela palavra `obj` no início e `endobj` no final. Esses objetos podem ser textos, desenhos, fontes, imagens, metadados, links, e um monte de coisas mais.
- Tabela de referência: uma lista de onde (no arquivo) está cada objeto. Esta tabela permite o acesso direto a cada um e a todos os objetos.
- Rodapé: Diversas informações que permitem o acesso rápido ao arquivo.

Nos interessa, neste exercício a informação `startxref` que é uma das últimas do arquivo. Você deve abrir o arquivo

```
arq09.pdf
```

publicado no lugar usual, com um editor de hexadecimal e deve procurar o startxref no final do arquivo. Na linha seguinte, existe um endereço que o PDF coloca em decimal. Você precisa transformar este número em hexadecimal e localizar no arquivo (mais acima) este endereço.

Nele, deve aparecer a palavra `xref` e na linha seguinte um zero e um número. Na linha seguinte, uma entrada (que todo arquivo tem) com muitos zeros, o número 65535 e uma letra. Na linha seguinte, **um número** seguido de zeros e uma letra. Trata-se do endereço do primeiro objeto do arquivo.

Responda aqui:

endereço da xref em hexadecimal	endereço do primeiro objeto em decimal

CEP - UP - UTFPR - PUCPr - UFPR –
17/06/2024 - 17:11:23.4
Prof Dr P Kantek (pkantek@gmail.com)
Objetos binários grandes vivxq10a, V: 1.01 76663
DANILO DEPETRIS SOARES 24CC2301 - 2
entregar ate 4/julho _____ / _____ / _____

Objetos binários grandes - BLOBs

O objetivo desta atividade e conhecer e manusear os chamados objetos binários grandes: imagens, vídeos, documentos, músicas, etc. Tais objetos quando estão na memória do computador são conhecidos como BLOBs e quando vão para uma memória não volátil (pendrives, discos, SSDs, nuvem, ...) passam a ser arquivos.

Resultados da digitalização crescente da nossa sociedade, precisam ser conhecidos para deles se extraírem todos os benefícios que eles apresentam.

Eis alguns raciocínios: a documentação em papel de um avião boeing 747 ocuparia 5 aviões 747 para ser transportada. Alguém ainda compra enciclopédias em papel ? Lembram de quando as fotografias eram analógicas ? Um filme Ektachrome (36 fotos) custava bem caro. Vou chutar uns 150 reais, o que daria quase 5 reais/foto, só o filme, sem contar a revelação e as ampliações. Quando o I.R. era em papel, a receita precisava montar uma operação de guerra, envolvendo todos os bancos, apenas para receber as declarações, e isso que era mais ou menos 1/3 do número de contribuintes atual. E as músicas: durante muitos anos, no dia do meu pagamento eu ia a uma loja de discos (?) e comprava 1 único LP, ($\pm$ 30 minutos de música) era o que o meu dinheiro alcançava. Aliás, falando de música, vale comentar a ascensão e queda do formato CD. Ele passou de maravilha tecnológica em meados dos 80 a algo obsoleto em meados dos 2010. Do céu ao inferno em 30 anos. E, se perder em uma cidade estranha, tendo apenas o guia 4 Rodas para ajudar ? Uma sensação indescritível.

Vamos estudar mais de perto alguns BLOBs, lembrando que qualquer um pode criar e chamar de seu, um BLOB. Pela sua definição (grande bloco de dados binários com uma lógica interna de organização e acesso) vê-se que isto é perfeitamente possível. Assim, vamos nos limitar a BLOBs padrão como formatos e conceitos conhecidos e publicados.

Imagem É difícil contar esta história. Ela começa no daguerrótipo de meados do século 19, através da fixação de imagens em placas de vidro. Segue pelas imagens de TV que eram capturadas e transmitidas eletronicamente. Seu registro foi feito em gravações magnéticas analógicas. Uma nova necessidade surgiu na obtenção de fotos na Lua e além, quando não havia como retornar filmes. A propósito, um pouco antes disto, a vigilância eletrônica na Guerra Fria era muito custosa: satélites fotografavam o inimigo e precisavam ejetar os filmes (que acabavam rápido) ao sobrevoar território amigo. A primeira foto da lua, transmitida eletronicamente foi feita pela espaçonave Ranger em 1964, 17 minutos antes dela se espatifar na superfície lunar. Quando a Internet começou (por volta de 1990) a necessidade de manusear imagens sofreu uma mudança: antes o gerador e o visualizador eram de mesma origem, pelo que não havia necessidade de padronização. Com a Internet essa história mudou. Uma das primeiras respostas a esta demanda foi o formato BMP (Bitmap Image File). Embora originalmente proposto pela Microsoft e IBM, teve seu padrão publicado e virou de fato, um formato livre. Sua origem é a bio-física dos olhos humanos (3 canais separados para Red, Green e Blue), além de um truque bem legal chamado mapeamento pelo qual se negocia tamanho do arquivo VERSUS menor quantidade de cores. A principal inovação do BMP foi criar, publicar e obedecer a um padrão. Parece pouco, mas não é. A chave da popularização e do sucesso sempre é um PADRãO. Anote isso na sua cabeça e nunca mais esqueça. O padrão pode ser ruim ou bom (neste caso não é muito), mas é um padrão. A imagem é uma matriz numérica (outra inovação entusiasmadora), logo sujeita a sofrer a ação maravilhosa de qualquer matemática.

Padrão BMP O arquivo começa com um bloco de controle de 54 bytes, que contém o identificador 'BM', o tamanho do objeto (arquivo), onde a imagem começa, L=largura e A=altura da imagem, profundidade). Depois, dependendo da profundidade, pode haver uma tabela de cores. Finalmente a imagem na forma de $L \times A$ se for uma imagem mapeada ou 3 tabelas de $L \times A$, uma para cada cor se for uma imagem true-color. Veja na imagem a seguir estes conceitos

```
Exemplo:
      0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
      -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
0000  42 4D 62 05 00 00 00 00 00 00 36 04 00 00 28 00 BM.....6...(
0010  00 00 12 00 00 00 0F 00 00 00 01 00 08 00 00 00 ................
0020  00 00 60 01 00 00 00 00 00 00 00 00 00 00 00 01 ..`.............
0030  00 00 00 00 00 00 00 00 FF FF 00 80 80 80 00 FF ................
0040  FF 00 FF 00 FF 00 00 00 00 FF 00 00 00 00 00 FF ................
0050  00 00 FF FF 00 00 FF 00 00 00 09 09 09 00 0A 0A ................
0060  0A 00 0B 0B 0B 00 0C 0C 0C 00 0D 0D 0D 00 0E 0E ................
0070  0E 00 0F 0F 0F 00 10 10 10 00 11 11 11 00 12 12 ................
...
0410  F6 00 F7 F7 F7 00 F8 F8 F8 00 F9 F9 F9 00 FA FA ................
0420  FA 00 FB FB FB 00 FC FC FC 00 FD FD FD 00 FE FE ................
0430  FE 00 FF FF FF 00 01 01 03 01 01 01 08 01 01 01 ................
0440  01 01 02 01 01 06 01 01 00 00 04 04 03 04 04 04 ................
0450  04 04 04 04 04 04 04 04 04 04 04 04 00 00 01 01 ................
```

Os problemas do BMP incluem: nenhuma compressão, o que graças ao fenômeno da *localidade de referência espacial* é uma tragédia. Também não prevê nenhum tipo de animação, o que faz falta na Internet. Não sabia (hoje sabe) lidar com transparência.

Um concorrente importante do BMP (lá nos primórdios) foi o formato GIF que admite compressão (usando o belo algoritmo LZW) e também admite uma forma rudimentar de animação. O problema do GIF é que ele foi e é um formato proprietário. Isto deu origem ao formato PNG que é praticamente o mesmo, mas livre.

A compressão poderia ser muito melhor, se se admitisse algum nivel de degradação na reconstituição da imagem após sua transformação em BLOB. Isto foi conseguido no padrão JPEG, que ao admitir degradações (sob controle), gera compressões muito maiores do que GIF ou similares. Outra vantagem é a possibilidade de criação em etapas das imagens (coisa que o BMP nunca ofereceu). O padrão foi formalizado pela ISO em 1991, mas na época disparou uma briga por patentes e direitos. Nos EUA, as últimas patentes expiraram em 2004. A compressão aqui é pelo uso da Transformada Discreta do Coseno. A sua operação está além do nosso interesse (querendo, olhe VIVX690c ou 690p), mas em resumo: a imagem é quebrada em blocos de 8×8; esta matriz é multiplicada pela matriz DCT, o que passa do domínio do pixel para o domínio da frequência (Transformada de Fourier); os coeficientes são quantizados; os coeficientes são comprimidos. Depois, se faz o inverso e *voilà*.

Som O MP3, sigla para MPEG-1 Audio Layer III, é um formato de arquivo de áudio digital que revolucionou a forma como consumimos e compartilhamos música.

Desenvolvido no início da década de 1990 pelo Moving Picture Experts Group (MPEG), o MP3 se tornou o formato dominante para música digital, utilizado em players de música portáteis, computadores, smartphones e serviços de streaming online. O MP3 utiliza uma técnica de compressão com perda chamada percepção auditiva psicoacústica para reduzir o tamanho dos arquivos de áudio. Essa técnica remove informações da faixa de áudio que o ouvido humano provavelmente não consegue perceber, resultando em arquivos menores sem comprometer significativamente a qualidade da música para a maioria dos ouvintes. Apesar da compressão, o MP3 oferece uma qualidade de som bastante boa, especialmente em taxas de bits mais altas (como 128 kbps ou superior). Tamanho de Arquivo Reduzido: Os arquivos MP3 são significativamente menores do que os arquivos de áudio não compactados, como CDs, permitindo fácil armazenamento e compartilhamento de música digital.

O MP3 teve um impacto profundo na indústria da música e na forma como consumimos música gerando um declínio das vendas de CDs: Os consumidores passaram a baixar e compartilhar músicas digitalmente. O MP3 impulsionou o crescimento da música digital, com o surgimento de serviços de streaming online como Spotify, Apple Music e Deezer. O MP3 abriu caminho para novos modelos de negócios na indústria da música, como a venda de músicas online e assinaturas de serviços de streaming.

Apesar de seu sucesso, o MP3 também enfrentou alguns desafios: Em taxas de bits mais baixas (como 64 kbps), a qualidade de som do MP3 pode ser perceptívelmente inferior à do áudio não compactado. Propriedades Intelectuais: Questões de direitos autorais e pirataria surgiram com a proliferação de arquivos MP3 compartilhados online. Formatos Alternativos: Novos formatos de áudio digital, como FLAC e AAC, oferecem maior qualidade de som ou taxas de compressão mais eficientes, mas ainda não alcançaram o mesmo nível de popularidade do MP3.

Outros padrões: MID (notação musical sintética), WAV (nenhuma perda de qualidade, logo arquivos grandes), além do formato OGG, da plataforma APPLE.

Vídeo No vídeo, estamos diante de necessidade maior de compressão. Se uma imagem estática já é grande, imagine precisar de 30 imagens estáticas por segundo. Agora, a localidade de referência ganha um novo viés, a l.r. no tempo. Ou seja, se imaginarmos um vídeo como um array 3-d (dimensões tempo, altura e largura), podemos comprimir em 2 dessas dimensões.

Aqui a história começa com os formatos analógicos (VHS, Betamax), segue pelo CD-ROM e VCD, e chega ao DVD. Depois vem os formatos AVC (padrão H.264), depois HEVC (H.265), VP9 (formato livre e aberto da Google) e AV1 (formato livre e aberto, melhor que o H.265). Diferentemente da imagem (e semelhantemente ao som), aqui existe muita preocupação com a latência (atraso no envio dos pacotes) que impacta diretamente a sincronia. O padrão H.264 utiliza timestamps e buffers para sincronizar áudio e vídeo. Em caso de perda de pacotes, mecanismos de recuperação de erros como o Concealment Motion Vector (CMV) podem ser utilizados para minimizar o impacto na sincronia. O CMV divide o vídeo em macroblocos. Quando há perda de um pacote contendo os vetores de movimento de um determinado macrobloco, o CMV busca os macroblocos vizinhos ($\uparrow, \downarrow, \leftarrow \rightarrow$) e ve se algum deles tem movimento similar ao esperado do macrobloco perdido. Daí ele "empresta" vetores e reconstitui o movimento ausente.

PDF O Portable Document Format (PDF), criado pela Adobe em 1993, se tornou um dos formatos de arquivo mais utilizados no mundo, presente em diversos setores e aplicações. Mais detalhes: vivx680a. Um documento PDF pode ser aberto e visualizado em qualquer dispositivo com um leitor de PDF instalado, independentemente do sistema operacional ou hardware utilizado. Isso garante que a formatação original do documento é sempre a mesma, independentemente da plataforma em que ele é aberto. Isso é crucial para documentos que exigem alta fidelidade visual, como apresentações, relatórios e publicações. O PDF oferece recursos de segurança robustos, como criptografia e assinaturas digitais, que protegem o conteúdo contra acessos não autorizados, modificações e falsificações. Isso torna o PDF ideal para documentos confidenciais e contratos legais. Até porque ele pode ser acompanhado de uma assinatura digital (*hash*). Suporta recursos de acessibilidade, como tags de estrutura e legendas para imagens, que facilitam o acesso ao conteúdo do documento por pessoas com deficiência visual ou outras necessidades especiais. Ao contrário de outros formatos de arquivo que podem ser corrompidos ou se tornar obsoletos com o tempo, o PDF é um formato durável e estável. Pense na economia de árvores: imagine se todos os PDFs existentes fossem para o papel. Ele é ideal para compartilhar documentos com outras pessoas, pois garante que a formatação original seja preservada e o conteúdo seja visualizado corretamente em diferentes dispositivos. O PDF é uma ótima opção para armazenar documentos importantes de forma segura e durável, pois protege o conteúdo contra modificações e garante que ele possa ser acessado no futuro. Revistas, livros, relatórios e outros materiais de publicação podem ser distribuídos no formato PDF, garantindo uma experiência de leitura consistente em diferentes dispositivos. **Outros:** Deixo de tratar aqui, por absoluta falta de espaço, outros BLOBs como exames médicos, arquivos de telescópios (vivx760), ...

Olhando dentro de um BLOB

Agora e hora de examinar com um "microscópio" as partes internas de um desses arquivos. O escolhido pela importância é o arquivo PDF. Pode parecer pouco, mas esta ideia simples, já salvou centenas de milhares (milhões ?) de árvores. Já são muitos os sistemas de informação que não geram mais saídas em papel, limitando-se a criar arquivos PDF.

Um arquivo PDF descreve como uma página ficará depois de impressa e ele é melhor do que imagem dessa página pelas seguintes razões

- ⋆ A imagem como tal sempre é muito maior do que sua descrição. A razão para isso é a evidente redundância de qualquer imagem.

⋆ A imagem como tal é fixa e só pode ser rotacionada, redimensionada ou de alguma maneira modificada à custa de processamento e perda de qualidade.

⋆ Sobre uma imagem não há como manipular o conteúdo, por exemplo conduzindo uma busca textual (^F alguma coisa ou então ^C e ^V).

⋆ Alterar uma única palavra sobre uma imagem exige habilidades de pintura. Sobre a descrição de uma página é tarefa simples: basta trocar uma palavra pela outra.

⋆ Imagine uma imagem enviada para um monitor (72 DPP) ou a mesma imagem enviada para uma impressora (300 DPP). Não podem ser a mesma imagem.

A versão 1.0 do padrão PDF nasceu em 1993 quando a Adobe lançou 2 programas: o Distiller (para gerar PDF) e o Reader (para ler): ambos pagos. A coisa começou a mudar quando o departamento de rendas americano comprou uma licença que permitia aos contribuintes baixar o Reader. Rapidamente a Adobe deu-se conta de que seria melhor para ela distribuir gratuitamente o Reader. Nos dez anos seguintes o PDF acabou se tornando o padrão de fato de descrição de imagens.

Apresenta as seguintes vantagens:

⋆ acesso aleatório: qualquer página pode ser montada independente das demais.

⋆ linearização: todos os elementos necessários para montar uma página estão juntos.

⋆ atualização incremental: mudanças ficam registradas no fim do arquivo. Vantagem: opção de desfazer ilimitada e rápida atualização. Desvantagem: o arquivo não pára de crescer.

⋆ fontes incorporadas: o destino sempre será montado corretamente. A fonte é desbastada de tudo que não é necessário.

⋆ padronização ISO: em 2008 a ISO abraçou o padrão Adobe PDF 1.7, o que o transformou em padrão aberto.

⋆ compatibilidade para trás e para a frente: Qualquer leitor lê as versões antigas e deve ler as futuras, já que todos os leitores ignoram o que não conhecem.

⋆ integrado a todos os principais browsers do protocolo `HTTP`.

Um arquivo PDF

Cabeçalho Formado por 2 linhas: a primeira identifica o arquivo como um PDF e informa a versão PDF em que ele foi gerado. A segunda linha inclui caracteres que não podem ser impressos (para avisar eventuais leitores deste fato). Exemplo:

```
%PDF-1.0
%âã
```

Corpo Um conjunto de nodos de um grafo, contendo as páginas, conteúdo gráfico, textos, sempre na forma de objetos. Cada objeto começa com um número de objeto (iniciando em 1), um número de geração (sempre 0) e a palavra chave `obj`. O objeto termina pela palavra chave `endobj`. Exemplo:

```
1 0 obj     -- descrevendo o objeto 1
<<          -- começa um dicionário
 /Kids [2 0 R]    -- /Kids tem o valor 2 0 R
 /Count 1         -- só um
 /Type /Pages     -- do tipo : página
>>               -- fim do dicionário
endobj            -- fim do objeto
```

Tabela de Referência Cruzada Lista o deslocamento em bytes de cada objeto no corpo do arquivo. Exemplo

```
0 6          -- seis entradas na tabela
0000000000  65535  f   -- entrada especial
0000000015  00000  n   -- obj1 no desloc 15
0000000074  00000  n   -- obj2 no desloc 74
...
```

Rodapé Informações e metadados para aumentar a performance de acesso. A primeira entrada é `trailer`. Depois o dicionário de trailer que deve apresentar ao menos a entrada `\Size` dizendo quantos itens há na tabela de referência cruzada e `\Root` dizendo qual objeto é o catálogo do documento. Depois vem uma linha que só contém `startxref` seguida por uma única linha indicando o deslocamento em bytes utilizado no inicio da tabela de referência cruzada do arquivo. Termina tudo a linha `%%EOF`. Exemplo:

```
trailer   -- palavra chave
<<   -- começa um dicionário
/Root 5 0 R   -- o catálogo é o objeto 5
/Size 6       -- tem o tamanho 6
>>   -- fim do dicionário
startxref
459  -- deslocamento da tabela de ref cruzada
%%EOF   -- fim de arquivo
```

Componentes

Uma descrição PDF suporta 5 componentes básicos: i. Números inteiros e reais (mas não exponenciais); ii. strings que são escritos entre parênteses; iii.Nomes, que sempre começam por uma barra normal; iv.Os valores booleanos `true` e `false` e v. O componente nulo (`null`). Aceita também 3 componentes compostos: i. Arrays, que são uma coleção ordenada de outros componentes, escritos entre colchetes; ii.Dicionários que são uma lista não ordenada de duplas: nome e componente. Começam por `<<` e terminam por `>>`. Finalmente iii. Fluxos, que armazenam dados binários, sempre acompanhados de um dicionário que descreve seus atributos, como comprimento e parâmetros de compactação, por exemplo.

☞ Para você fazer

As coisas em um arquivo PDF são separadas por um divisor de linhas. Pode ser o caracter $X'10'$, ou o $X'13'$ ou uma combinação de ambos. Ao examinar um PDF em hexadecimal, acostume-se a separar as linhas marcando este caractere. Isto posto, os componentes principais do arquivo PDF são:

- Header: identificação PDF e uma coleção de letras acentuadas
- Corpo: objetos numerados a partir de 1 sequencialmente, e identificados pela palavra `obj` no início e `endobj` no final. Esses objetos podem ser textos, desenhos, fontes, imagens, metadados, links, e um monte de coisas mais.
- Tabela de referência: uma lista de onde (no arquivo) está cada objeto. Esta tabela permite o acesso direto a cada um e a todos os objetos.
- Rodapé: Diversas informações que permitem o acesso rápido ao arquivo.

Nos interessa, neste exercício a informação `startxref` que é uma das últimas do arquivo. Você deve abrir o arquivo

```
arq10.pdf
```

publicado no lugar usual, com um editor de hexadecimal e deve procurar o startxref no final do arquivo. Na linha seguinte, existe um endereço que o PDF coloca em decimal. Você precisa transformar este número em hexadecimal e localizar no arquivo (mais acima) este endereço.

Nele, deve aparecer a palavra `xref` e na linha seguinte um zero e um número. Na linha seguinte, uma entrada (que todo arquivo tem) com muitos zeros, o número 65535 e uma letra. Na linha seguinte, **um número** seguido de zeros e uma letra. Trata-se do endereço do primeiro objeto do arquivo.

Responda aqui:

endereço da xref em hexadecimal	endereço do primeiro objeto em decimal

CEP - UP - UTFPR - PUCPr - UFPR –
17/06/2024 - 17:11:23.4
Prof Dr P Kantek (pkantek@gmail.com)
Objetos binários grandes vivxq10a, V: 1.01 76513
EDUARDO CAVALHEIRO DIAS 24CC2301 - 3
entregar ate 4/julho _____ / _____ / _____

Objetos binários grandes - BLOBs

O objetivo desta atividade e conhecer e manusear os chamados objetos binários grandes: imagens, vídeos, documentos, músicas, etc. Tais objetos quando estão na memória do computador são conhecidos como BLOBs e quando vão para uma memória não volátil (pendrives, discos, SSDs, nuvem, ...) passam a ser arquivos.

Resultados da digitalização crescente da nossa sociedade, precisam ser conhecidos para deles se extraírem todos os benefícios que eles apresentam.

Eis alguns raciocínios: a documentação em papel de um avião boeing 747 ocuparia 5 aviões 747 para ser transportada. Alguém ainda compra enciclopédias em papel ? Lembram de quando as fotografias eram analógicas ? Um filme Ektachrome (36 fotos) custava bem caro. Vou chutar uns 150 reais, o que daria quase 5 reais/foto, só o filme, sem contar a revelação e as ampliações. Quando o I.R. era em papel, a receita precisava montar uma operação de guerra, envolvendo todos os bancos, apenas para receber as declarações, e isso que era mais ou menos 1/3 do número de contribuintes atual. E as músicas: durante muitos anos, no dia do meu pagamento eu ia a uma loja de discos (?) e comprava 1 único LP, ($\pm$ 30 minutos de música) era o que o meu dinheiro alcançava. Aliás, falando de música, vale comentar a ascensão e queda do formato CD. Ele passou de maravilha tecnológica em meados dos 80 a algo obsoleto em meados dos 2010. Do céu ao inferno em 30 anos. E, se perder em uma cidade estranha, tendo apenas o guia 4 Rodas para ajudar ? Uma sensação indescritível.

Vamos estudar mais de perto alguns BLOBs, lembrando que qualquer um pode criar e chamar de seu, um BLOB. Pela sua definição (grande bloco de dados binários com uma lógica interna de organização e acesso) vê-se que isto é perfeitamente possível. Assim, vamos nos limitar a BLOBs padrão como formatos e conceitos conhecidos e publicados.

Imagem É difícil contar esta história. Ela começa no daguerrótipo de meados do século 19, através da fixação de imagens em placas de vidro. Segue pelas imagens de TV que eram capturadas e transmitidas eletronicamente. Seu registro foi feito em gravações magnéticas analógicas. Uma nova necessidade surgiu na obtenção de fotos na Lua e além, quando não havia como retornar filmes. A propósito, um pouco antes disto, a vigilância eletrônica na Guerra Fria era muito custosa: satélites fotografavam o inimigo e precisavam ejetar os filmes (que acabavam rápido) ao sobrevoar território amigo. A primeira foto da lua, transmitida eletronicamente foi feita pela espaçonave Ranger em 1964, 17 minutos antes dela se espatifar na superfície lunar. Quando a Internet começou (por volta de 1990) a necessidade de manusear imagens sofreu uma mudança: antes o gerador e o visualizador eram de mesma origem, pelo que não havia necessidade de padronização. Com a Internet essa história mudou. Uma das primeiras respostas a esta demanda foi o formato BMP (Bitmap Image File). Embora originalmente proposto pela Microsoft e IBM, teve seu padrão publicado e virou de fato, um formato livre. Sua origem é a bio-física dos olhos humanos (3 canais separados para Red, Green e Blue), além de um truque bem legal chamado mapeamento pelo qual se negocia tamanho do arquivo VERSUS menor quantidade de cores. A principal inovação do BMP foi criar, publicar e obedecer a um padrão. Parece pouco, mas não é. A chave da popularização e do sucesso sempre é um PADRãO. Anote isso na sua cabeça e nunca mais esqueça. O padrão pode ser ruim ou bom (neste caso não é muito), mas é um padrão. A imagem é uma matriz numérica (outra inovação entusiasmadora), logo sujeita a sofrer a ação maravilhosa de qualquer matemática.

Padrão BMP O arquivo começa com um bloco de controle de 54 bytes, que contém o identificador 'BM', o tamanho do objeto (arquivo), onde a imagem começa, L=largura e A=altura da imagem, profundidade). Depois, dependendo da profundidade, pode haver uma tabela de cores. Finalmente a imagem na forma de $L \times A$ se for uma imagem mapeada ou 3 tabelas de $L \times A$, uma para cada cor se for uma imagem true-color. Veja na imagem a seguir estes conceitos

Exemplo:

```
      0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
      -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
0000  42 4D 62 05 00 00 00 00 00 00 36 04 00 00 28 00 BM.....6...(
0010  00 00 12 00 00 00 0F 00 00 00 01 00 08 00 00 00 ................
0020  00 00 60 01 00 00 00 00 00 00 00 00 00 00 00 00 ..`.............
0030  00 00 00 00 00 00 00 00 FF FF 00 80 80 80 00 FF ................
0040  FF 00 FF 00 FF 00 00 00 00 FF 00 00 00 00 00 00 ................
0050  00 00 FF FF 00 00 FF 00 00 00 09 09 09 00 0A 0A ................
0060  0A 00 0B 0B 0B 00 0C 0C 0C 00 0D 0D 0D 00 0E 0E ................
0070  0E 00 0F 0F 0F 00 10 10 10 00 11 11 11 00 12 12 ................
...
0410  F6 00 F7 F7 F7 00 F8 F8 F8 00 F9 F9 F9 00 FA FA ................
0420  FA 00 FB FB FB 00 FC FC FC 00 FD FD FD 00 FE FE ................
0430  FE 00 FF FF FF 00 01 01 03 01 01 01 08 01 01 01 ................
0440  01 01 02 01 01 06 01 01 00 00 04 04 03 04 04 04 ................
0450  04 04 04 04 04 04 04 04 04 04 04 04 00 00 01 01 ................
```

Os problemas do BMP incluem: nenhuma compressão, o que graças ao fenômeno da *localidade de referência espacial* é uma tragédia. Também não prevê nenhum tipo de animação, o que faz falta na Internet. Não sabia (hoje sabe) lidar com transparência.

Um concorrente importante do BMP (lá nos primórdios) foi o formato GIF que admite compressão (usando o belo algoritmo LZW) e também admite uma forma rudimentar de animação. O problema do GIF é que ele foi e é um formato proprietário. Isto deu origem ao formato PNG que é praticamente o mesmo, mas livre.

A compressão poderia ser muito melhor, se se admitisse algum nivel de degradação na reconstituição da imagem após sua transformação em BLOB. Isto foi conseguido no padrão JPEG, que ao admitir degradações (sob controle), gera compressões muito maiores do que GIF ou similares. Outra vantagem é a possibilidade de criação em etapas das imagens (coisa que o BMP nunca ofereceu). O padrão foi formalizado pela ISO em 1991, mas na época disparou uma briga por patentes e direitos. Nos EUA, as últimas patentes expiraram em 2004. A compressão aqui é pelo uso da Transformada Discreta do Coseno. A sua operação está além do nosso interesse (querendo, olhe VIVX690c ou 690p), mas em resumo: a imagem é quebrada em blocos de 8×8; esta matriz é multiplicada pela matriz DCT, o que passa do domínio do pixel para o domínio da frequência (Transformada de Fourier); os coeficientes são quantizados; os coeficientes são comprimidos. Depois, se faz o inverso e *voilà*.

Som O MP3, sigla para MPEG-1 Audio Layer III, é um formato de arquivo de áudio digital que revolucionou a forma como consumimos e compartilhamos música.

Desenvolvido no início da década de 1990 pelo Moving Picture Experts Group (MPEG), o MP3 se tornou o formato dominante para música digital, utilizado em players de música portáteis, computadores, smartphones e serviços de streaming online. O MP3 utiliza uma técnica de compressão com perda chamada percepção auditiva psicoacústica para reduzir o tamanho dos arquivos de áudio. Essa técnica remove informações da faixa de áudio que o ouvido humano provavelmente não consegue perceber, resultando em arquivos menores sem comprometer significativamente a qualidade da música para a maioria dos ouvintes. Apesar da compressão, o MP3 oferece uma qualidade de som bastante boa, especialmente em taxas de bits mais altas (como 128 kbps ou superior). Tamanho de Arquivo Reduzido: Os arquivos MP3 são significativamente menores do que os arquivos de áudio não compactados, como CDs, permitindo fácil armazenamento e compartilhamento de música digital.

O MP3 teve um impacto profundo na indústria da música e na forma como consumimos música gerando um declínio das vendas de CDs: Os consumidores passaram a baixar e compartilhar músicas digitalmente. O MP3 impulsionou o crescimento da música digital, com o surgimento de serviços de streaming online como Spotify, Apple Music e Deezer. O MP3 abriu caminho para novos modelos de negócios na indústria da música, como a venda de músicas online e assinaturas de serviços de streaming.

Apesar de seu sucesso, o MP3 também enfrentou alguns desafios: Em taxas de bits mais baixas (como 64 kbps), a qualidade de som do MP3 pode ser perceptívelmente inferior à do áudio não compactado. Propriedades Intelectuais: Questões de direitos autorais e pirataria surgiram com a proliferação de arquivos MP3 compartilhados online. Formatos Alternativos: Novos formatos de áudio digital, como FLAC e AAC, oferecem maior qualidade de som ou taxas de compressão mais eficientes, mas ainda não alcançaram o mesmo nível de popularidade do MP3.

Outros padrões: MID (notação musical sintética), WAV (nenhuma perda de qualidade, logo arquivos grandes), além do formato OGG, da plataforma APPLE.

Vídeo No vídeo, estamos diante de necessidade maior de compressão. Se uma imagem estática já é grande, imagine precisar de 30 imagens estáticas por segundo. Agora, a localidade de referência ganha um novo viés, a l.r. no tempo. Ou seja, se imaginarmos um vídeo como um array 3-d (dimensões tempo, altura e largura), podemos comprimir em 2 dessas dimensões.

Aqui a história começa com os formatos analógicos (VHS, Betamax), segue pelo CD-ROM e VCD, e chega ao DVD. Depois vem os formatos AVC (padrão H.264), depois HEVC (H.265), VP9 (formato livre e aberto da Google) e AV1 (formato livre e aberto, melhor que o H.265). Diferentemente da imagem (e semelhantemente ao som), aqui existe muita preocupação com a latência (atraso no envio dos pacotes) que impacta diretamente a sincronia. O padrão H.264 utiliza timestamps e buffers para sincronizar áudio e vídeo. Em caso de perda de pacotes, mecanismos de recuperação de erros como o Concealment Motion Vector (CMV) podem ser utilizados para minimizar o impacto na sincronia. O CMV divide o vídeo em macroblocos. Quando há perda de um pacote contendo os vetores de movimento de um determinado macrobloco, o CMV busca os macroblocos vizinhos ($\uparrow, \downarrow, \leftarrow \rightarrow$) e ve se algum deles tem movimento similar ao esperado do macrobloco perdido. Daí ele "empresta" vetores e reconstitui o movimento ausente.

PDF O Portable Document Format (PDF), criado pela Adobe em 1993, se tornou um dos formatos de arquivo mais utilizados no mundo, presente em diversos setores e aplicações. Mais detalhes: vivx680a. Um documento PDF pode ser aberto e visualizado em qualquer dispositivo com um leitor de PDF instalado, independentemente do sistema operacional ou hardware utilizado. Isso garante que a formatação original do documento é sempre a mesma, independentemente da plataforma em que ele é aberto. Isso é crucial para documentos que exigem alta fidelidade visual, como apresentações, relatórios e publicações. O PDF oferece recursos de segurança robustos, como criptografia e assinaturas digitais, que protegem o conteúdo contra acessos não autorizados, modificações e falsificações. Isso torna o PDF ideal para documentos confidenciais e contratos legais. Até porque ele pode ser acompanhado de uma assinatura digital (*hash*). Suporta recursos de acessibilidade, como tags de estrutura e legendas para imagens, que facilitam o acesso ao conteúdo do documento por pessoas com deficiência visual ou outras necessidades especiais. Ao contrário de outros formatos de arquivo que podem ser corrompidos ou se tornar obsoletos com o tempo, o PDF é um formato durável e estável. Pense na economia de árvores: imagine se todos os PDFs existentes fossem para o papel. Ele é ideal para compartilhar documentos com outras pessoas, pois garante que a formatação original seja preservada e o conteúdo seja visualizado corretamente em diferentes dispositivos. O PDF é uma ótima opção para armazenar documentos importantes de forma segura e durável, pois protege o conteúdo contra modificações e garante que ele possa ser acessado no futuro. Revistas, livros, relatórios e outros materiais de publicação podem ser distribuídos no formato PDF, garantindo uma experiência de leitura consistente em diferentes dispositivos. **Outros:** Deixo de tratar aqui, por absoluta falta de espaço, outros BLOBs como exames médicos, arquivos de telescópios (vivx760), ...

Olhando dentro de um BLOB

Agora e hora de examinar com um "microscópio" as partes internas de um desses arquivos. O escolhido pela importância é o arquivo PDF. Pode parecer pouco, mas esta ideia simples, já salvou centenas de milhares (milhões ?) de árvores. Já são muitos os sistemas de informação que não geram mais saídas em papel, limitando-se a criar arquivos PDF.

Um arquivo PDF descreve como uma página ficará depois de impressa e ele é melhor do que imagem dessa página pelas seguintes razões

- ⋆ A imagem como tal sempre é muito maior do que sua descrição. A razão para isso é a evidente redundância de qualquer imagem.

⋆ A imagem como tal é fixa e só pode ser rotacionada, redimensionada ou de alguma maneira modificada à custa de processamento e perda de qualidade.

⋆ Sobre uma imagem não há como manipular o conteúdo, por exemplo conduzindo uma busca textual (^F alguma coisa ou então ^C e ^V).

⋆ Alterar uma única palavra sobre uma imagem exige habilidades de pintura. Sobre a descrição de uma página é tarefa simples: basta trocar uma palavra pela outra.

⋆ Imagine uma imagem enviada para um monitor (72 DPP) ou a mesma imagem enviada para uma impressora (300 DPP). Não podem ser a mesma imagem.

A versão 1.0 do padrão PDF nasceu em 1993 quando a Adobe lançou 2 programas: o Distiller (para gerar PDF) e o Reader (para ler): ambos pagos. A coisa começou a mudar quando o departamento de rendas americano comprou uma licença que permitia aos contribuintes baixar o Reader. Rapidamente a Adobe deu-se conta de que seria melhor para ela distribuir gratuitamente o Reader. Nos dez anos seguintes o PDF acabou se tornando o padrão de fato de descrição de imagens.

Apresenta as seguintes vantagens:

⋆ acesso aleatório: qualquer página pode ser montada independente das demais.

⋆ linearização: todos os elementos necessários para montar uma página estão juntos.

⋆ atualização incremental: mudanças ficam registradas no fim do arquivo. Vantagem: opção de desfazer ilimitada e rápida atualização. Desvantagem: o arquivo não pára de crescer.

⋆ fontes incorporadas: o destino sempre será montado corretamente. A fonte é desbastada de tudo que não é necessário.

⋆ padronização ISO: em 2008 a ISO abraçou o padrão Adobe PDF 1.7, o que o transformou em padrão aberto.

⋆ compatibilidade para trás e para a frente: Qualquer leitor lê as versões antigas e deve ler as futuras, já que todos os leitores ignoram o que não conhecem.

⋆ integrado a todos os principais browsers do protocolo `HTTP`.

Um arquivo PDF

Cabeçalho Formado por 2 linhas: a primeira identifica o arquivo como um PDF e informa a versão PDF em que ele foi gerado. A segunda linha inclui caracteres que não podem ser impressos (para avisar eventuais leitores deste fato). Exemplo:

```
%PDF-1.0
%âã
```

Corpo Um conjunto de nodos de um grafo, contendo as páginas, conteúdo gráfico, textos, sempre na forma de objetos. Cada objeto começa com um número de objeto (iniciando em 1), um número de geração (sempre 0) e a palavra chave `obj`. O objeto termina pela palavra chave `endobj`. Exemplo:

```
1 0 obj     -- descrevendo o objeto 1
<<          -- começa um dicionário
 /Kids [2 0 R]     -- /Kids tem o valor 2 0 R
 /Count 1          -- só um
 /Type /Pages      -- do tipo : página
>>                 -- fim do dicionário
endobj             -- fim do objeto
```

Tabela de Referência Cruzada Lista o deslocamento em bytes de cada objeto no corpo do arquivo. Exemplo

```
0 6           -- seis entradas na tabela
0000000000  65535  f   -- entrada especial
0000000015  00000  n   -- obj1 no desloc 15
0000000074  00000  n   -- obj2 no desloc 74
...
```

Rodapé Informações e metadados para aumentar a performance de acesso. A primeira entrada é `trailer`. Depois o dicionário de trailer que deve apresentar ao menos a entrada `\Size` dizendo quantos itens há na tabela de referência cruzada e `\Root` dizendo qual objeto é o catálogo do documento. Depois vem uma linha que só contém `startxref` seguida por uma única linha indicando o deslocamento em bytes utilizado no inicio da tabela de referência cruzada do arquivo. Termina tudo a linha `%%EOF`. Exemplo:

```
trailer   -- palavra chave
<<   -- começa um dicionário
/Root 5 0 R   -- o catálogo é o objeto 5
/Size 6       -- tem o tamanho 6
>>   -- fim do dicionário
startxref
459  -- deslocamento da tabela de ref cruzada
%%EOF   -- fim de arquivo
```

Componentes

Uma descrição PDF suporta 5 componentes básicos: i. Números inteiros e reais (mas não exponenciais); ii. strings que são escritos entre parênteses; iii.Nomes, que sempre começam por uma barra normal; iv.Os valores booleanos `true` e `false` e v. O componente nulo (`null`). Aceita também 3 componentes compostos: i. Arrays, que são uma coleção ordenada de outros componentes, escritos entre colchetes; ii.Dicionários que são uma lista não ordenada de duplas: nome e componente. Começam por `<<` e terminam por `>>`. Finalmente iii. Fluxos, que armazenam dados binários, sempre acompanhados de um dicionário que descreve seus atributos, como comprimento e parâmetros de compactação, por exemplo.

☞ Para você fazer

As coisas em um arquivo PDF são separadas por um divisor de linhas. Pode ser o caracter $X'10'$, ou o $X'13'$ ou uma combinação de ambos. Ao examinar um PDF em hexadecimal, acostume-se a separar as linhas marcando este caractere. Isto posto, os componentes principais do arquivo PDF são:

- Header: identificação PDF e uma coleção de letras acentuadas
- Corpo: objetos numerados a partir de 1 sequencialmente, e identificados pela palavra `obj` no início e `endobj` no final. Esses objetos podem ser textos, desenhos, fontes, imagens, metadados, links, e um monte de coisas mais.
- Tabela de referência: uma lista de onde (no arquivo) está cada objeto. Esta tabela permite o acesso direto a cada um e a todos os objetos.
- Rodapé: Diversas informações que permitem o acesso rápido ao arquivo.

Nos interessa, neste exercício a informação `startxref` que é uma das últimas do arquivo. Você deve abrir o arquivo

```
arq11.pdf
```

publicado no lugar usual, com um editor de hexadecimal e deve procurar o startxref no final do arquivo. Na linha seguinte, existe um endereço que o PDF coloca em decimal. Você precisa transformar este número em hexadecimal e localizar no arquivo (mais acima) este endereço.

Nele, deve aparecer a palavra `xref` e na linha seguinte um zero e um número. Na linha seguinte, uma entrada (que todo arquivo tem) com muitos zeros, o número 65535 e uma letra. Na linha seguinte, **um número** seguido de zeros e uma letra. Trata-se do endereço do primeiro objeto do arquivo.

Responda aqui:

endereço da xref em hexadecimal	endereço do primeiro objeto em decimal

CEP - UP - UTFPR - PUCPr - UFPR –
17/06/2024 - 17:11:23.4
Prof Dr P Kantek (pkantek@gmail.com)
Objetos binários grandes vivxq10a, V: 1.01 76520
ERIC CORDEIRO 24CC2301 - 4 entregar ate 4/julho _____ / _____ / _____

Objetos binários grandes - BLOBs

O objetivo desta atividade e conhecer e manusear os chamados objetos binários grandes: imagens, vídeos, documentos, músicas, etc. Tais objetos quando estão na memória do computador são conhecidos como BLOBs e quando vão para uma memória não volátil (pendrives, discos, SSDs, nuvem, ...) passam a ser arquivos.

Resultados da digitalização crescente da nossa sociedade, precisam ser conhecidos para deles se extraírem todos os benefícios que eles apresentam.

Eis alguns raciocínios: a documentação em papel de um avião boeing 747 ocuparia 5 aviões 747 para ser transportada. Alguém ainda compra enciclopédias em papel ? Lembram de quando as fotografias eram analógicas ? Um filme Ektachrome (36 fotos) custava bem caro. Vou chutar uns 150 reais, o que daria quase 5 reais/foto, só o filme, sem contar a revelação e as ampliações. Quando o I.R. era em papel, a receita precisava montar uma operação de guerra, envolvendo todos os bancos, apenas para receber as declarações, e isso que era mais ou menos 1/3 do número de contribuintes atual. E as músicas: durante muitos anos, no dia do meu pagamento eu ia a uma loja de discos (?) e comprava 1 único LP, ($\pm$ 30 minutos de música) era o que o meu dinheiro alcançava. Aliás, falando de música, vale comentar a ascensão e queda do formato CD. Ele passou de maravilha tecnológica em meados dos 80 a algo obsoleto em meados dos 2010. Do céu ao inferno em 30 anos. E, se perder em uma cidade estranha, tendo apenas o guia 4 Rodas para ajudar ? Uma sensação indescritível.

Vamos estudar mais de perto alguns BLOBs, lembrando que qualquer um pode criar e chamar de seu, um BLOB. Pela sua definição (grande bloco de dados binários com uma lógica interna de organização e acesso) vê-se que isto é perfeitamente possível. Assim, vamos nos limitar a BLOBs padrão como formatos e conceitos conhecidos e publicados.

Imagem É difícil contar esta história. Ela começa no daguerrótipo de meados do século 19, através da fixação de imagens em placas de vidro. Segue pelas imagens de TV que eram capturadas e transmitidas eletronicamente. Seu registro foi feito em gravações magnéticas analógicas. Uma nova necessidade surgiu na obtenção de fotos na Lua e além, quando não havia como retornar filmes. A propósito, um pouco antes disto, a vigilância eletrônica na Guerra Fria era muito custosa: satélites fotografavam o inimigo e precisavam ejetar os filmes (que acabavam rápido) ao sobrevoar território amigo. A primeira foto da lua, transmitida eletronicamente foi feita pela espaçonave Ranger em 1964, 17 minutos antes dela se espatifar na superfície lunar. Quando a Internet começou (por volta de 1990) a necessidade de manusear imagens sofreu uma mudança: antes o gerador e o visualizador eram de mesma origem, pelo que não havia necessidade de padronização. Com a Internet essa história mudou. Uma das primeiras respostas a esta demanda foi o formato BMP (Bitmap Image File). Embora originalmente proposto pela Microsoft e IBM, teve seu padrão publicado e virou de fato, um formato livre. Sua origem é a bio-física dos olhos humanos (3 canais separados para Red, Green e Blue), além de um truque bem legal chamado mapeamento pelo qual se negocia tamanho do arquivo VERSUS menor quantidade de cores. A principal inovação do BMP foi criar, publicar e obedecer a um padrão. Parece pouco, mas não é. A chave da popularização e do sucesso sempre é um PADRãO. Anote isso na sua cabeça e nunca mais esqueça. O padrão pode ser ruim ou bom (neste caso não é muito), mas é um padrão. A imagem é uma matriz numérica (outra inovação entusiasmadora), logo sujeita a sofrer a ação maravilhosa de qualquer matemática.

Padrão BMP O arquivo começa com um bloco de controle de 54 bytes, que contém o identificador 'BM', o tamanho do objeto (arquivo), onde a imagem começa, L=largura e A=altura da imagem, profundidade). Depois, dependendo da profundidade, pode haver uma tabela de cores. Finalmente a imagem na forma de $L \times A$ se for uma imagem mapeada ou 3 tabelas de $L \times A$, uma para cada cor se for uma imagem true-color. Veja na imagem a seguir estes conceitos

```
Exemplo:
      0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
      -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
0000  42 4D 62 05 00 00 00 00 00 00 36 04 00 00 28 00 BM.....6...(
0010  00 00 12 00 00 00 0F 00 00 00 01 00 08 00 00 00 ................
0020  00 00 60 01 00 00 00 00 00 00 00 00 00 00 00 01 ..`.............
0030  00 00 00 00 00 00 00 00 FF FF 00 80 80 80 00 FF ................
0040  FF 00 FF 00 FF 00 00 00 00 FF 00 00 00 00 00 FF ................
0050  00 00 FF FF 00 00 FF 00 00 00 09 09 09 00 0A 0A ................
0060  0A 00 0B 0B 0B 00 0C 0C 0C 00 0D 0D 0D 00 0E 0E ................
0070  0E 00 0F 0F 0F 00 10 10 10 00 11 11 11 00 12 12 ................
...
0410  F6 00 F7 F7 F7 00 F8 F8 F8 00 F9 F9 F9 00 FA FA ................
0420  FA 00 FB FB FB 00 FC FC FC 00 FD FD FD 00 FE FE ................
0430  FE 00 FF FF FF 00 01 01 03 01 01 01 08 01 01 01 ................
0440  01 01 02 01 01 06 01 01 00 00 04 04 03 04 04 04 ................
0450  04 04 04 04 04 04 04 04 04 04 04 04 00 00 01 01 ................
```

Os problemas do BMP incluem: nenhuma compressão, o que graças ao fenômeno da *localidade de referência espacial* é uma tragédia. Também não prevê nenhum tipo de animação, o que faz falta na Internet. Não sabia (hoje sabe) lidar com transparência.

Um concorrente importante do BMP (lá nos primórdios) foi o formato GIF que admite compressão (usando o belo algoritmo LZW) e também admite uma forma rudimentar de animação. O problema do GIF é que ele foi e é um formato proprietário. Isto deu origem ao formato PNG que é praticamente o mesmo, mas livre.

A compressão poderia ser muito melhor, se se admitisse algum nivel de degradação na reconstituição da imagem após sua transformação em BLOB. Isto foi conseguido no padrão JPEG, que ao admitir degradações (sob controle), gera compressões muito maiores do que GIF ou similares. Outra vantagem é a possibilidade de criação em etapas das imagens (coisa que o BMP nunca ofereceu). O padrão foi formalizado pela ISO em 1991, mas na época disparou uma briga por patentes e direitos. Nos EUA, as últimas patentes expiraram em 2004. A compressão aqui é pelo uso da Transformada Discreta do Coseno. A sua operação está além do nosso interesse (querendo, olhe VIVX690c ou 690p), mas em resumo: a imagem é quebrada em blocos de 8×8; esta matriz é multiplicada pela matriz DCT, o que passa do domínio do pixel para o domínio da frequência (Transformada de Fourier); os coeficientes são quantizados; os coeficientes são comprimidos. Depois, se faz o inverso e *voilà*.

Som O MP3, sigla para MPEG-1 Audio Layer III, é um formato de arquivo de áudio digital que revolucionou a forma como consumimos e compartilhamos música.

Desenvolvido no início da década de 1990 pelo Moving Picture Experts Group (MPEG), o MP3 se tornou o formato dominante para música digital, utilizado em players de música portáteis, computadores, smartphones e serviços de streaming online. O MP3 utiliza uma técnica de compressão com perda chamada percepção auditiva psicoacústica para reduzir o tamanho dos arquivos de áudio. Essa técnica remove informações da faixa de áudio que o ouvido humano provavelmente não consegue perceber, resultando em arquivos menores sem comprometer significativamente a qualidade da música para a maioria dos ouvintes. Apesar da compressão, o MP3 oferece uma qualidade de som bastante boa, especialmente em taxas de bits mais altas (como 128 kbps ou superior). Tamanho de Arquivo Reduzido: Os arquivos MP3 são significativamente menores do que os arquivos de áudio não compactados, como CDs, permitindo fácil armazenamento e compartilhamento de música digital.

O MP3 teve um impacto profundo na indústria da música e na forma como consumimos música gerando um declínio das vendas de CDs: Os consumidores passaram a baixar e compartilhar músicas digitalmente. O MP3 impulsionou o crescimento da música digital, com o surgimento de serviços de streaming online como Spotify, Apple Music e Deezer. O MP3 abriu caminho para novos modelos de negócios na indústria da música, como a venda de músicas online e assinaturas de serviços de streaming.

Apesar de seu sucesso, o MP3 também enfrentou alguns desafios: Em taxas de bits mais baixas (como 64 kbps), a qualidade de som do MP3 pode ser perceptívelmente inferior à do áudio não compactado. Propriedades Intelectuais: Questões de direitos autorais e pirataria surgiram com a proliferação de arquivos MP3 compartilhados online. Formatos Alternativos: Novos formatos de áudio digital, como FLAC e AAC, oferecem maior qualidade de som ou taxas de compressão mais eficientes, mas ainda não alcançaram o mesmo nível de popularidade do MP3.

Outros padrões: MID (notação musical sintética), WAV (nenhuma perda de qualidade, logo arquivos grandes), além do formato OGG, da plataforma APPLE.

Vídeo No vídeo, estamos diante de necessidade maior de compressão. Se uma imagem estática já é grande, imagine precisar de 30 imagens estáticas por segundo. Agora, a localidade de referência ganha um novo viés, a l.r. no tempo. Ou seja, se imaginarmos um vídeo como um array 3-d (dimensões tempo, altura e largura), podemos comprimir em 2 dessas dimensões.

Aqui a história começa com os formatos analógicos (VHS, Betamax), segue pelo CD-ROM e VCD, e chega ao DVD. Depois vem os formatos AVC (padrão H.264), depois HEVC (H.265), VP9 (formato livre e aberto da Google) e AV1 (formato livre e aberto, melhor que o H.265). Diferentemente da imagem (e semelhantemente ao som), aqui existe muita preocupação com a latência (atraso no envio dos pacotes) que impacta diretamente a sincronia. O padrão H.264 utiliza timestamps e buffers para sincronizar áudio e vídeo. Em caso de perda de pacotes, mecanismos de recuperação de erros como o Concealment Motion Vector (CMV) podem ser utilizados para minimizar o impacto na sincronia. O CMV divide o vídeo em macroblocos. Quando há perda de um pacote contendo os vetores de movimento de um determinado macrobloco, o CMV busca os macroblocos vizinhos ($\uparrow, \downarrow, \leftarrow \rightarrow$) e ve se algum deles tem movimento similar ao esperado do macrobloco perdido. Daí ele "empresta" vetores e reconstitui o movimento ausente.

PDF O Portable Document Format (PDF), criado pela Adobe em 1993, se tornou um dos formatos de arquivo mais utilizados no mundo, presente em diversos setores e aplicações. Mais detalhes: vivx680a. Um documento PDF pode ser aberto e visualizado em qualquer dispositivo com um leitor de PDF instalado, independentemente do sistema operacional ou hardware utilizado. Isso garante que a formatação original do documento é sempre a mesma, independentemente da plataforma em que ele é aberto. Isso é crucial para documentos que exigem alta fidelidade visual, como apresentações, relatórios e publicações. O PDF oferece recursos de segurança robustos, como criptografia e assinaturas digitais, que protegem o conteúdo contra acessos não autorizados, modificações e falsificações. Isso torna o PDF ideal para documentos confidenciais e contratos legais. Até porque ele pode ser acompanhado de uma assinatura digital (*hash*). Suporta recursos de acessibilidade, como tags de estrutura e legendas para imagens, que facilitam o acesso ao conteúdo do documento por pessoas com deficiência visual ou outras necessidades especiais. Ao contrário de outros formatos de arquivo que podem ser corrompidos ou se tornar obsoletos com o tempo, o PDF é um formato durável e estável. Pense na economia de árvores: imagine se todos os PDFs existentes fossem para o papel. Ele é ideal para compartilhar documentos com outras pessoas, pois garante que a formatação original seja preservada e o conteúdo seja visualizado corretamente em diferentes dispositivos. O PDF é uma ótima opção para armazenar documentos importantes de forma segura e durável, pois protege o conteúdo contra modificações e garante que ele possa ser acessado no futuro. Revistas, livros, relatórios e outros materiais de publicação podem ser distribuídos no formato PDF, garantindo uma experiência de leitura consistente em diferentes dispositivos. **Outros:** Deixo de tratar aqui, por absoluta falta de espaço, outros BLOBs como exames médicos, arquivos de telescópios (vivx760), ...

Olhando dentro de um BLOB

Agora e hora de examinar com um "microscópio" as partes internas de um desses arquivos. O escolhido pela importância é o arquivo PDF. Pode parecer pouco, mas esta ideia simples, já salvou centenas de milhares (milhões ?) de árvores. Já são muitos os sistemas de informação que não geram mais saídas em papel, limitando-se a criar arquivos PDF.

Um arquivo PDF descreve como uma página ficará depois de impressa e ele é melhor do que imagem dessa página pelas seguintes razões

- ⋆ A imagem como tal sempre é muito maior do que sua descrição. A razão para isso é a evidente redundância de qualquer imagem.

⋆ A imagem como tal é fixa e só pode ser rotacionada, redimensionada ou de alguma maneira modificada à custa de processamento e perda de qualidade.

⋆ Sobre uma imagem não há como manipular o conteúdo, por exemplo conduzindo uma busca textual (^F alguma coisa ou então ^C e ^V).

⋆ Alterar uma única palavra sobre uma imagem exige habilidades de pintura. Sobre a descrição de uma página é tarefa simples: basta trocar uma palavra pela outra.

⋆ Imagine uma imagem enviada para um monitor (72 DPP) ou a mesma imagem enviada para uma impressora (300 DPP). Não podem ser a mesma imagem.

A versão 1.0 do padrão PDF nasceu em 1993 quando a Adobe lançou 2 programas: o Distiller (para gerar PDF) e o Reader (para ler): ambos pagos. A coisa começou a mudar quando o departamento de rendas americano comprou uma licença que permitia aos contribuintes baixar o Reader. Rapidamente a Adobe deu-se conta de que seria melhor para ela distribuir gratuitamente o Reader. Nos dez anos seguintes o PDF acabou se tornando o padrão de fato de descrição de imagens.

Apresenta as seguintes vantagens:

⋆ acesso aleatório: qualquer página pode ser montada independente das demais.

⋆ linearização: todos os elementos necessários para montar uma página estão juntos.

⋆ atualização incremental: mudanças ficam registradas no fim do arquivo. Vantagem: opção de desfazer ilimitada e rápida atualização. Desvantagem: o arquivo não pára de crescer.

⋆ fontes incorporadas: o destino sempre será montado corretamente. A fonte é desbastada de tudo que não é necessário.

⋆ padronização ISO: em 2008 a ISO abraçou o padrão Adobe PDF 1.7, o que o transformou em padrão aberto.

⋆ compatibilidade para trás e para a frente: Qualquer leitor lê as versões antigas e deve ler as futuras, já que todos os leitores ignoram o que não conhecem.

⋆ integrado a todos os principais browsers do protocolo `HTTP`.

Um arquivo PDF

Cabeçalho Formado por 2 linhas: a primeira identifica o arquivo como um PDF e informa a versão PDF em que ele foi gerado. A segunda linha inclui caracteres que não podem ser impressos (para avisar eventuais leitores deste fato). Exemplo:

```
%PDF-1.0
%âã
```

Corpo Um conjunto de nodos de um grafo, contendo as páginas, conteúdo gráfico, textos, sempre na forma de objetos. Cada objeto começa com um número de objeto (iniciando em 1), um número de geração (sempre 0) e a palavra chave `obj`. O objeto termina pela palavra chave `endobj`. Exemplo:

```
1 0 obj     -- descrevendo o objeto 1
<<          -- começa um dicionário
 /Kids [2 0 R]    -- /Kids tem o valor 2 0 R
 /Count 1         -- só um
 /Type /Pages     -- do tipo : página
>>               -- fim do dicionário
endobj            -- fim do objeto
```

Tabela de Referência Cruzada Lista o deslocamento em bytes de cada objeto no corpo do arquivo. Exemplo

```
0 6          -- seis entradas na tabela
0000000000  65535  f   -- entrada especial
0000000015  00000  n   -- obj1 no desloc 15
0000000074  00000  n   -- obj2 no desloc 74
...
```

Rodapé Informações e metadados para aumentar a performance de acesso. A primeira entrada é `trailer`. Depois o dicionário de trailer que deve apresentar ao menos a entrada `\Size` dizendo quantos itens há na tabela de referência cruzada e `\Root` dizendo qual objeto é o catálogo do documento. Depois vem uma linha que só contém `startxref` seguida por uma única linha indicando o deslocamento em bytes utilizado no inicio da tabela de referência cruzada do arquivo. Termina tudo a linha `%%EOF`. Exemplo:

```
trailer   -- palavra chave
<<   -- começa um dicionário
/Root 5 0 R   -- o catálogo é o objeto 5
/Size 6       -- tem o tamanho 6
>>   -- fim do dicionário
startxref
459  -- deslocamento da tabela de ref cruzada
%%EOF   -- fim de arquivo
```

Componentes

Uma descrição PDF suporta 5 componentes básicos: i. Números inteiros e reais (mas não exponenciais); ii. strings que são escritos entre parênteses; iii.Nomes, que sempre começam por uma barra normal; iv.Os valores booleanos `true` e `false` e v. O componente nulo (`null`). Aceita também 3 componentes compostos: i. Arrays, que são uma coleção ordenada de outros componentes, escritos entre colchetes; ii.Dicionários que são uma lista não ordenada de duplas: nome e componente. Começam por `<<` e terminam por `>>`. Finalmente iii. Fluxos, que armazenam dados binários, sempre acompanhados de um dicionário que descreve seus atributos, como comprimento e parâmetros de compactação, por exemplo.

☞ Para você fazer

As coisas em um arquivo PDF são separadas por um divisor de linhas. Pode ser o caracter $X'10'$, ou o $X'13'$ ou uma combinação de ambos. Ao examinar um PDF em hexadecimal, acostume-se a separar as linhas marcando este caractere. Isto posto, os componentes principais do arquivo PDF são:

- Header: identificação PDF e uma coleção de letras acentuadas
- Corpo: objetos numerados a partir de 1 sequencialmente, e identificados pela palavra `obj` no início e `endobj` no final. Esses objetos podem ser textos, desenhos, fontes, imagens, metadados, links, e um monte de coisas mais.
- Tabela de referência: uma lista de onde (no arquivo) está cada objeto. Esta tabela permite o acesso direto a cada um e a todos os objetos.
- Rodapé: Diversas informações que permitem o acesso rápido ao arquivo.

Nos interessa, neste exercício a informação `startxref` que é uma das últimas do arquivo. Você deve abrir o arquivo

```
arq05.pdf
```

publicado no lugar usual, com um editor de hexadecimal e deve procurar o startxref no final do arquivo. Na linha seguinte, existe um endereço que o PDF coloca em decimal. Você precisa transformar este número em hexadecimal e localizar no arquivo (mais acima) este endereço.

Nele, deve aparecer a palavra `xref` e na linha seguinte um zero e um número. Na linha seguinte, uma entrada (que todo arquivo tem) com muitos zeros, o número 65535 e uma letra. Na linha seguinte, **um número** seguido de zeros e uma letra. Trata-se do endereço do primeiro objeto do arquivo.

Responda aqui:

endereço da xref em hexadecimal	endereço do primeiro objeto em decimal

CEP - UP - UTFPR - PUCPr - UFPR –
17/06/2024 - 17:11:23.4
Prof Dr P Kantek (pkantek@gmail.com)
Objetos binários grandes vivxq10a, V: 1.01 76537
FELIPE EIJI KANO 24CC2301 - 5 entregar ate 4/julho _____ / _____ / _____

Objetos binários grandes - BLOBs

O objetivo desta atividade e conhecer e manusear os chamados objetos binários grandes: imagens, vídeos, documentos, músicas, etc. Tais objetos quando estão na memória do computador são conhecidos como BLOBs e quando vão para uma memória não volátil (pendrives, discos, SSDs, nuvem, ...) passam a ser arquivos.

Resultados da digitalização crescente da nossa sociedade, precisam ser conhecidos para deles se extraírem todos os benefícios que eles apresentam.

Eis alguns raciocínios: a documentação em papel de um avião boeing 747 ocuparia 5 aviões 747 para ser transportada. Alguém ainda compra enciclopédias em papel ? Lembram de quando as fotografias eram analógicas ? Um filme Ektachrome (36 fotos) custava bem caro. Vou chutar uns 150 reais, o que daria quase 5 reais/foto, só o filme, sem contar a revelação e as ampliações. Quando o I.R. era em papel, a receita precisava montar uma operação de guerra, envolvendo todos os bancos, apenas para receber as declarações, e isso que era mais ou menos 1/3 do número de contribuintes atual. E as músicas: durante muitos anos, no dia do meu pagamento eu ia a uma loja de discos (?) e comprava 1 único LP, ($\pm$ 30 minutos de música) era o que o meu dinheiro alcançava. Aliás, falando de música, vale comentar a ascensão e queda do formato CD. Ele passou de maravilha tecnológica em meados dos 80 a algo obsoleto em meados dos 2010. Do céu ao inferno em 30 anos. E, se perder em uma cidade estranha, tendo apenas o guia 4 Rodas para ajudar ? Uma sensação indescritível.

Vamos estudar mais de perto alguns BLOBs, lembrando que qualquer um pode criar e chamar de seu, um BLOB. Pela sua definição (grande bloco de dados binários com uma lógica interna de organização e acesso) vê-se que isto é perfeitamente possível. Assim, vamos nos limitar a BLOBs padrão como formatos e conceitos conhecidos e publicados.

Imagem É difícil contar esta história. Ela começa no daguerrótipo de meados do século 19, através da fixação de imagens em placas de vidro. Segue pelas imagens de TV que eram capturadas e transmitidas eletronicamente. Seu registro foi feito em gravações magnéticas analógicas. Uma nova necessidade surgiu na obtenção de fotos na Lua e além, quando não havia como retornar filmes. A propósito, um pouco antes disto, a vigilância eletrônica na Guerra Fria era muito custosa: satélites fotografavam o inimigo e precisavam ejetar os filmes (que acabavam rápido) ao sobrevoar território amigo. A primeira foto da lua, transmitida eletronicamente foi feita pela espaçonave Ranger em 1964, 17 minutos antes dela se espatifar na superfície lunar. Quando a Internet começou (por volta de 1990) a necessidade de manusear imagens sofreu uma mudança: antes o gerador e o visualizador eram de mesma origem, pelo que não havia necessidade de padronização. Com a Internet essa história mudou. Uma das primeiras respostas a esta demanda foi o formato BMP (Bitmap Image File). Embora originalmente proposto pela Microsoft e IBM, teve seu padrão publicado e virou de fato, um formato livre. Sua origem é a bio-física dos olhos humanos (3 canais separados para Red, Green e Blue), além de um truque bem legal chamado mapeamento pelo qual se negocia tamanho do arquivo VERSUS menor quantidade de cores. A principal inovação do BMP foi criar, publicar e obedecer a um padrão. Parece pouco, mas não é. A chave da popularização e do sucesso sempre é um PADRãO. Anote isso na sua cabeça e nunca mais esqueça. O padrão pode ser ruim ou bom (neste caso não é muito), mas é um padrão. A imagem é uma matriz numérica (outra inovação entusiasmadora), logo sujeita a sofrer a ação maravilhosa de qualquer matemática.

Padrão BMP O arquivo começa com um bloco de controle de 54 bytes, que contém o identificador 'BM', o tamanho do objeto (arquivo), onde a imagem começa, L=largura e A=altura da imagem, profundidade). Depois, dependendo da profundidade, pode haver uma tabela de cores. Finalmente a imagem na forma de $L \times A$ se for uma imagem mapeada ou 3 tabelas de $L \times A$, uma para cada cor se for uma imagem true-color. Veja na imagem a seguir estes conceitos

Exemplo:

```
      0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
      -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
0000  42 4D 62 05 00 00 00 00 00 00 36 04 00 00 28 00 BM.....6...(
0010  00 00 12 00 00 00 0F 00 00 00 01 00 08 00 00 00 ................
0020  00 00 60 01 00 00 00 00 00 00 00 00 00 00 00 00 ..`.............
0030  00 00 00 00 00 00 00 00 FF FF 00 80 80 80 00 FF ................
0040  FF 00 FF 00 FF 00 00 00 00 FF 00 00 00 00 00 00 ................
0050  00 00 FF FF 00 00 FF 00 00 00 09 09 09 00 0A 0A ................
0060  0A 00 0B 0B 0B 00 0C 0C 0C 00 0D 0D 0D 00 0E 0E ................
0070  0E 00 0F 0F 0F 00 10 10 10 00 11 11 11 00 12 12 ................
...
0410  F6 00 F7 F7 F7 00 F8 F8 F8 00 F9 F9 F9 00 FA FA ................
0420  FA 00 FB FB FB 00 FC FC FC 00 FD FD FD 00 FE FE ................
0430  FE 00 FF FF FF 00 01 01 03 01 01 01 08 01 01 01 ................
0440  01 01 02 01 01 06 01 01 00 00 04 04 03 04 04 04 ................
0450  04 04 04 04 04 04 04 04 04 04 04 04 00 00 01 01 ................
```

Os problemas do BMP incluem: nenhuma compressão, o que graças ao fenômeno da *localidade de referência espacial* é uma tragédia. Também não prevê nenhum tipo de animação, o que faz falta na Internet. Não sabia (hoje sabe) lidar com transparência.

Um concorrente importante do BMP (lá nos primórdios) foi o formato GIF que admite compressão (usando o belo algoritmo LZW) e também admite uma forma rudimentar de animação. O problema do GIF é que ele foi e é um formato proprietário. Isto deu origem ao formato PNG que é praticamente o mesmo, mas livre.

A compressão poderia ser muito melhor, se se admitisse algum nivel de degradação na reconstituição da imagem após sua transformação em BLOB. Isto foi conseguido no padrão JPEG, que ao admitir degradações (sob controle), gera compressões muito maiores do que GIF ou similares. Outra vantagem é a possibilidade de criação em etapas das imagens (coisa que o BMP nunca ofereceu). O padrão foi formalizado pela ISO em 1991, mas na época disparou uma briga por patentes e direitos. Nos EUA, as últimas patentes expiraram em 2004. A compressão aqui é pelo uso da Transformada Discreta do Coseno. A sua operação está além do nosso interesse (querendo, olhe VIVX690c ou 690p), mas em resumo: a imagem é quebrada em blocos de 8×8; esta matriz é multiplicada pela matriz DCT, o que passa do domínio do pixel para o domínio da frequência (Transformada de Fourier); os coeficientes são quantizados; os coeficientes são comprimidos. Depois, se faz o inverso e *voilà*.

Som O MP3, sigla para MPEG-1 Audio Layer III, é um formato de arquivo de áudio digital que revolucionou a forma como consumimos e compartilhamos música.

Desenvolvido no início da década de 1990 pelo Moving Picture Experts Group (MPEG), o MP3 se tornou o formato dominante para música digital, utilizado em players de música portáteis, computadores, smartphones e serviços de streaming online. O MP3 utiliza uma técnica de compressão com perda chamada percepção auditiva psicoacústica para reduzir o tamanho dos arquivos de áudio. Essa técnica remove informações da faixa de áudio que o ouvido humano provavelmente não consegue perceber, resultando em arquivos menores sem comprometer significativamente a qualidade da música para a maioria dos ouvintes. Apesar da compressão, o MP3 oferece uma qualidade de som bastante boa, especialmente em taxas de bits mais altas (como 128 kbps ou superior). Tamanho de Arquivo Reduzido: Os arquivos MP3 são significativamente menores do que os arquivos de áudio não compactados, como CDs, permitindo fácil armazenamento e compartilhamento de música digital.

O MP3 teve um impacto profundo na indústria da música e na forma como consumimos música gerando um declínio das vendas de CDs: Os consumidores passaram a baixar e compartilhar músicas digitalmente. O MP3 impulsionou o crescimento da música digital, com o surgimento de serviços de streaming online como Spotify, Apple Music e Deezer. O MP3 abriu caminho para novos modelos de negócios na indústria da música, como a venda de músicas online e assinaturas de serviços de streaming.

Apesar de seu sucesso, o MP3 também enfrentou alguns desafios: Em taxas de bits mais baixas (como 64 kbps), a qualidade de som do MP3 pode ser perceptívelmente inferior à do áudio não compactado. Propriedades Intelectuais: Questões de direitos autorais e pirataria surgiram com a proliferação de arquivos MP3 compartilhados online. Formatos Alternativos: Novos formatos de áudio digital, como FLAC e AAC, oferecem maior qualidade de som ou taxas de compressão mais eficientes, mas ainda não alcançaram o mesmo nível de popularidade do MP3.

Outros padrões: MID (notação musical sintética), WAV (nenhuma perda de qualidade, logo arquivos grandes), além do formato OGG, da plataforma APPLE.

Vídeo No vídeo, estamos diante de necessidade maior de compressão. Se uma imagem estática já é grande, imagine precisar de 30 imagens estáticas por segundo. Agora, a localidade de referência ganha um novo viés, a l.r. no tempo. Ou seja, se imaginarmos um vídeo como um array 3-d (dimensões tempo, altura e largura), podemos comprimir em 2 dessas dimensões.

Aqui a história começa com os formatos analógicos (VHS, Betamax), segue pelo CD-ROM e VCD, e chega ao DVD. Depois vem os formatos AVC (padrão H.264), depois HEVC (H.265), VP9 (formato livre e aberto da Google) e AV1 (formato livre e aberto, melhor que o H.265). Diferentemente da imagem (e semelhantemente ao som), aqui existe muita preocupação com a latência (atraso no envio dos pacotes) que impacta diretamente a sincronia. O padrão H.264 utiliza timestamps e buffers para sincronizar áudio e vídeo. Em caso de perda de pacotes, mecanismos de recuperação de erros como o Concealment Motion Vector (CMV) podem ser utilizados para minimizar o impacto na sincronia. O CMV divide o vídeo em macroblocos. Quando há perda de um pacote contendo os vetores de movimento de um determinado macrobloco, o CMV busca os macroblocos vizinhos ($\uparrow, \downarrow, \leftarrow \rightarrow$) e ve se algum deles tem movimento similar ao esperado do macrobloco perdido. Daí ele "empresta" vetores e reconstitui o movimento ausente.

PDF O Portable Document Format (PDF), criado pela Adobe em 1993, se tornou um dos formatos de arquivo mais utilizados no mundo, presente em diversos setores e aplicações. Mais detalhes: vivx680a. Um documento PDF pode ser aberto e visualizado em qualquer dispositivo com um leitor de PDF instalado, independentemente do sistema operacional ou hardware utilizado. Isso garante que a formatação original do documento é sempre a mesma, independentemente da plataforma em que ele é aberto. Isso é crucial para documentos que exigem alta fidelidade visual, como apresentações, relatórios e publicações. O PDF oferece recursos de segurança robustos, como criptografia e assinaturas digitais, que protegem o conteúdo contra acessos não autorizados, modificações e falsificações. Isso torna o PDF ideal para documentos confidenciais e contratos legais. Até porque ele pode ser acompanhado de uma assinatura digital (*hash*). Suporta recursos de acessibilidade, como tags de estrutura e legendas para imagens, que facilitam o acesso ao conteúdo do documento por pessoas com deficiência visual ou outras necessidades especiais. Ao contrário de outros formatos de arquivo que podem ser corrompidos ou se tornar obsoletos com o tempo, o PDF é um formato durável e estável. Pense na economia de árvores: imagine se todos os PDFs existentes fossem para o papel. Ele é ideal para compartilhar documentos com outras pessoas, pois garante que a formatação original seja preservada e o conteúdo seja visualizado corretamente em diferentes dispositivos. O PDF é uma ótima opção para armazenar documentos importantes de forma segura e durável, pois protege o conteúdo contra modificações e garante que ele possa ser acessado no futuro. Revistas, livros, relatórios e outros materiais de publicação podem ser distribuídos no formato PDF, garantindo uma experiência de leitura consistente em diferentes dispositivos. **Outros:** Deixo de tratar aqui, por absoluta falta de espaço, outros BLOBs como exames médicos, arquivos de telescópios (vivx760), ...

Olhando dentro de um BLOB

Agora e hora de examinar com um "microscópio" as partes internas de um desses arquivos. O escolhido pela importância é o arquivo PDF. Pode parecer pouco, mas esta ideia simples, já salvou centenas de milhares (milhões ?) de árvores. Já são muitos os sistemas de informação que não geram mais saídas em papel, limitando-se a criar arquivos PDF.

Um arquivo PDF descreve como uma página ficará depois de impressa e ele é melhor do que imagem dessa página pelas seguintes razões

- ⋆ A imagem como tal sempre é muito maior do que sua descrição. A razão para isso é a evidente redundância de qualquer imagem.

⋆ A imagem como tal é fixa e só pode ser rotacionada, redimensionada ou de alguma maneira modificada à custa de processamento e perda de qualidade.

⋆ Sobre uma imagem não há como manipular o conteúdo, por exemplo conduzindo uma busca textual (^F alguma coisa ou então ^C e ^V).

⋆ Alterar uma única palavra sobre uma imagem exige habilidades de pintura. Sobre a descrição de uma página é tarefa simples: basta trocar uma palavra pela outra.

⋆ Imagine uma imagem enviada para um monitor (72 DPP) ou a mesma imagem enviada para uma impressora (300 DPP). Não podem ser a mesma imagem.

A versão 1.0 do padrão PDF nasceu em 1993 quando a Adobe lançou 2 programas: o Distiller (para gerar PDF) e o Reader (para ler): ambos pagos. A coisa começou a mudar quando o departamento de rendas americano comprou uma licença que permitia aos contribuintes baixar o Reader. Rapidamente a Adobe deu-se conta de que seria melhor para ela distribuir gratuitamente o Reader. Nos dez anos seguintes o PDF acabou se tornando o padrão de fato de descrição de imagens.

Apresenta as seguintes vantagens:

⋆ acesso aleatório: qualquer página pode ser montada independente das demais.

⋆ linearização: todos os elementos necessários para montar uma página estão juntos.

⋆ atualização incremental: mudanças ficam registradas no fim do arquivo. Vantagem: opção de desfazer ilimitada e rápida atualização. Desvantagem: o arquivo não pára de crescer.

⋆ fontes incorporadas: o destino sempre será montado corretamente. A fonte é desbastada de tudo que não é necessário.

⋆ padronização ISO: em 2008 a ISO abraçou o padrão Adobe PDF 1.7, o que o transformou em padrão aberto.

⋆ compatibilidade para trás e para a frente: Qualquer leitor lê as versões antigas e deve ler as futuras, já que todos os leitores ignoram o que não conhecem.

⋆ integrado a todos os principais browsers do protocolo `HTTP`.

Um arquivo PDF

Cabeçalho Formado por 2 linhas: a primeira identifica o arquivo como um PDF e informa a versão PDF em que ele foi gerado. A segunda linha inclui caracteres que não podem ser impressos (para avisar eventuais leitores deste fato). Exemplo:

```
%PDF-1.0
%âã
```

Corpo Um conjunto de nodos de um grafo, contendo as páginas, conteúdo gráfico, textos, sempre na forma de objetos. Cada objeto começa com um número de objeto (iniciando em 1), um número de geração (sempre 0) e a palavra chave `obj`. O objeto termina pela palavra chave `endobj`. Exemplo:

```
1 0 obj     -- descrevendo o objeto 1
<<          -- começa um dicionário
 /Kids [2 0 R]    -- /Kids tem o valor 2 0 R
 /Count 1         -- só um
 /Type /Pages     -- do tipo : página
>>               -- fim do dicionário
endobj            -- fim do objeto
```

Tabela de Referência Cruzada Lista o deslocamento em bytes de cada objeto no corpo do arquivo. Exemplo

```
0 6          -- seis entradas na tabela
0000000000  65535  f   -- entrada especial
0000000015  00000  n   -- obj1 no desloc 15
0000000074  00000  n   -- obj2 no desloc 74
...
```

Rodapé Informações e metadados para aumentar a performance de acesso. A primeira entrada é `trailer`. Depois o dicionário de trailer que deve apresentar ao menos a entrada `\Size` dizendo quantos itens há na tabela de referência cruzada e `\Root` dizendo qual objeto é o catálogo do documento. Depois vem uma linha que só contém `startxref` seguida por uma única linha indicando o deslocamento em bytes utilizado no inicio da tabela de referência cruzada do arquivo. Termina tudo a linha `%%EOF`. Exemplo:

```
trailer   -- palavra chave
<<   -- começa um dicionário
/Root 5 0 R   -- o catálogo é o objeto 5
/Size 6       -- tem o tamanho 6
>>   -- fim do dicionário
startxref
459  -- deslocamento da tabela de ref cruzada
%%EOF   -- fim de arquivo
```

Componentes

Uma descrição PDF suporta 5 componentes básicos: i. Números inteiros e reais (mas não exponenciais); ii. strings que são escritos entre parênteses; iii.Nomes, que sempre começam por uma barra normal; iv.Os valores booleanos `true` e `false` e v. O componente nulo (`null`). Aceita também 3 componentes compostos: i. Arrays, que são uma coleção ordenada de outros componentes, escritos entre colchetes; ii.Dicionários que são uma lista não ordenada de duplas: nome e componente. Começam por `<<` e terminam por `>>`. Finalmente iii. Fluxos, que armazenam dados binários, sempre acompanhados de um dicionário que descreve seus atributos, como comprimento e parâmetros de compactação, por exemplo.

☞ Para você fazer

As coisas em um arquivo PDF são separadas por um divisor de linhas. Pode ser o caracter $X'10'$, ou o $X'13'$ ou uma combinação de ambos. Ao examinar um PDF em hexadecimal, acostume-se a separar as linhas marcando este caractere. Isto posto, os componentes principais do arquivo PDF são:

- Header: identificação PDF e uma coleção de letras acentuadas
- Corpo: objetos numerados a partir de 1 sequencialmente, e identificados pela palavra `obj` no início e `endobj` no final. Esses objetos podem ser textos, desenhos, fontes, imagens, metadados, links, e um monte de coisas mais.
- Tabela de referência: uma lista de onde (no arquivo) está cada objeto. Esta tabela permite o acesso direto a cada um e a todos os objetos.
- Rodapé: Diversas informações que permitem o acesso rápido ao arquivo.

Nos interessa, neste exercício a informação `startxref` que é uma das últimas do arquivo. Você deve abrir o arquivo

```
arq03.pdf
```

publicado no lugar usual, com um editor de hexadecimal e deve procurar o startxref no final do arquivo. Na linha seguinte, existe um endereço que o PDF coloca em decimal. Você precisa transformar este número em hexadecimal e localizar no arquivo (mais acima) este endereço.

Nele, deve aparecer a palavra `xref` e na linha seguinte um zero e um número. Na linha seguinte, uma entrada (que todo arquivo tem) com muitos zeros, o número 65535 e uma letra. Na linha seguinte, **um número** seguido de zeros e uma letra. Trata-se do endereço do primeiro objeto do arquivo.

Responda aqui:

endereço da xref em hexadecimal	endereço do primeiro objeto em decimal

CEP - UP - UTFPR - PUCPr - UFPR –
17/06/2024 - 17:11:23.4
Prof Dr P Kantek (pkantek@gmail.com)
Objetos binários grandes vivxq10a, V: 1.01 76544
FELIPE JESMIEL LEITE 24CC2301 - 6 entregar ate 4/julho _____ / _____ / _____

Objetos binários grandes - BLOBs

O objetivo desta atividade e conhecer e manusear os chamados objetos binários grandes: imagens, vídeos, documentos, músicas, etc. Tais objetos quando estão na memória do computador são conhecidos como BLOBs e quando vão para uma memória não volátil (pendrives, discos, SSDs, nuvem, ...) passam a ser arquivos.

Resultados da digitalização crescente da nossa sociedade, precisam ser conhecidos para deles se extraírem todos os benefícios que eles apresentam.

Eis alguns raciocínios: a documentação em papel de um avião boeing 747 ocuparia 5 aviões 747 para ser transportada. Alguém ainda compra enciclopédias em papel ? Lembram de quando as fotografias eram analógicas ? Um filme Ektachrome (36 fotos) custava bem caro. Vou chutar uns 150 reais, o que daria quase 5 reais/foto, só o filme, sem contar a revelação e as ampliações. Quando o I.R. era em papel, a receita precisava montar uma operação de guerra, envolvendo todos os bancos, apenas para receber as declarações, e isso que era mais ou menos 1/3 do número de contribuintes atual. E as músicas: durante muitos anos, no dia do meu pagamento eu ia a uma loja de discos (?) e comprava 1 único LP, ($\pm$ 30 minutos de música) era o que o meu dinheiro alcançava. Aliás, falando de música, vale comentar a ascensão e queda do formato CD. Ele passou de maravilha tecnológica em meados dos 80 a algo obsoleto em meados dos 2010. Do céu ao inferno em 30 anos. E, se perder em uma cidade estranha, tendo apenas o guia 4 Rodas para ajudar ? Uma sensação indescritível.

Vamos estudar mais de perto alguns BLOBs, lembrando que qualquer um pode criar e chamar de seu, um BLOB. Pela sua definição (grande bloco de dados binários com uma lógica interna de organização e acesso) vê-se que isto é perfeitamente possível. Assim, vamos nos limitar a BLOBs padrão como formatos e conceitos conhecidos e publicados.

Imagem É difícil contar esta história. Ela começa no daguerrótipo de meados do século 19, através da fixação de imagens em placas de vidro. Segue pelas imagens de TV que eram capturadas e transmitidas eletronicamente. Seu registro foi feito em gravações magnéticas analógicas. Uma nova necessidade surgiu na obtenção de fotos na Lua e além, quando não havia como retornar filmes. A propósito, um pouco antes disto, a vigilância eletrônica na Guerra Fria era muito custosa: satélites fotografavam o inimigo e precisavam ejetar os filmes (que acabavam rápido) ao sobrevoar território amigo. A primeira foto da lua, transmitida eletronicamente foi feita pela espaçonave Ranger em 1964, 17 minutos antes dela se espatifar na superfície lunar. Quando a Internet começou (por volta de 1990) a necessidade de manusear imagens sofreu uma mudança: antes o gerador e o visualizador eram de mesma origem, pelo que não havia necessidade de padronização. Com a Internet essa história mudou. Uma das primeiras respostas a esta demanda foi o formato BMP (Bitmap Image File). Embora originalmente proposto pela Microsoft e IBM, teve seu padrão publicado e virou de fato, um formato livre. Sua origem é a bio-física dos olhos humanos (3 canais separados para Red, Green e Blue), além de um truque bem legal chamado mapeamento pelo qual se negocia tamanho do arquivo VERSUS menor quantidade de cores. A principal inovação do BMP foi criar, publicar e obedecer a um padrão. Parece pouco, mas não é. A chave da popularização e do sucesso sempre é um PADRãO. Anote isso na sua cabeça e nunca mais esqueça. O padrão pode ser ruim ou bom (neste caso não é muito), mas é um padrão. A imagem é uma matriz numérica (outra inovação entusiasmadora), logo sujeita a sofrer a ação maravilhosa de qualquer matemática.

Padrão BMP O arquivo começa com um bloco de controle de 54 bytes, que contém o identificador 'BM', o tamanho do objeto (arquivo), onde a imagem começa, L=largura e A=altura da imagem, profundidade). Depois, dependendo da profundidade, pode haver uma tabela de cores. Finalmente a imagem na forma de $L \times A$ se for uma imagem mapeada ou 3 tabelas de $L \times A$, uma para cada cor se for uma imagem true-color. Veja na imagem a seguir estes conceitos

```
Exemplo:
      0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
      -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
0000  42 4D 62 05 00 00 00 00 00 00 36 04 00 00 28 00 BM.....6...(
0010  00 00 12 00 00 00 0F 00 00 00 01 00 08 00 00 00 ................
0020  00 00 60 01 00 00 00 00 00 00 00 00 00 00 00 01 ..`.............
0030  00 00 00 00 00 00 00 00 FF FF 00 80 80 80 00 FF ................
0040  FF 00 FF 00 FF 00 00 00 00 FF 00 00 00 00 00 FF ................
0050  00 00 FF FF 00 00 FF 00 00 00 09 09 09 00 0A 0A ................
0060  0A 00 0B 0B 0B 00 0C 0C 0C 00 0D 0D 0D 00 0E 0E ................
0070  0E 00 0F 0F 0F 00 10 10 10 00 11 11 11 00 12 12 ................
...
0410  F6 00 F7 F7 F7 00 F8 F8 F8 00 F9 F9 F9 00 FA FA ................
0420  FA 00 FB FB FB 00 FC FC FC 00 FD FD FD 00 FE FE ................
0430  FE 00 FF FF FF 00 01 01 03 01 01 01 08 01 01 01 ................
0440  01 01 02 01 01 06 01 01 00 00 04 04 03 04 04 04 ................
0450  04 04 04 04 04 04 04 04 04 04 04 04 00 00 01 01 ................
```

Os problemas do BMP incluem: nenhuma compressão, o que graças ao fenômeno da *localidade de referência espacial* é uma tragédia. Também não prevê nenhum tipo de animação, o que faz falta na Internet. Não sabia (hoje sabe) lidar com transparência.

Um concorrente importante do BMP (lá nos primórdios) foi o formato GIF que admite compressão (usando o belo algoritmo LZW) e também admite uma forma rudimentar de animação. O problema do GIF é que ele foi e é um formato proprietário. Isto deu origem ao formato PNG que é praticamente o mesmo, mas livre.

A compressão poderia ser muito melhor, se se admitisse algum nivel de degradação na reconstituição da imagem após sua transformação em BLOB. Isto foi conseguido no padrão JPEG, que ao admitir degradações (sob controle), gera compressões muito maiores do que GIF ou similares. Outra vantagem é a possibilidade de criação em etapas das imagens (coisa que o BMP nunca ofereceu). O padrão foi formalizado pela ISO em 1991, mas na época disparou uma briga por patentes e direitos. Nos EUA, as últimas patentes expiraram em 2004. A compressão aqui é pelo uso da Transformada Discreta do Coseno. A sua operação está além do nosso interesse (querendo, olhe VIVX690c ou 690p), mas em resumo: a imagem é quebrada em blocos de 8×8; esta matriz é multiplicada pela matriz DCT, o que passa do domínio do pixel para o domínio da frequência (Transformada de Fourier); os coeficientes são quantizados; os coeficientes são comprimidos. Depois, se faz o inverso e *voilà*.

Som O MP3, sigla para MPEG-1 Audio Layer III, é um formato de arquivo de áudio digital que revolucionou a forma como consumimos e compartilhamos música.

Desenvolvido no início da década de 1990 pelo Moving Picture Experts Group (MPEG), o MP3 se tornou o formato dominante para música digital, utilizado em players de música portáteis, computadores, smartphones e serviços de streaming online. O MP3 utiliza uma técnica de compressão com perda chamada percepção auditiva psicoacústica para reduzir o tamanho dos arquivos de áudio. Essa técnica remove informações da faixa de áudio que o ouvido humano provavelmente não consegue perceber, resultando em arquivos menores sem comprometer significativamente a qualidade da música para a maioria dos ouvintes. Apesar da compressão, o MP3 oferece uma qualidade de som bastante boa, especialmente em taxas de bits mais altas (como 128 kbps ou superior). Tamanho de Arquivo Reduzido: Os arquivos MP3 são significativamente menores do que os arquivos de áudio não compactados, como CDs, permitindo fácil armazenamento e compartilhamento de música digital.

O MP3 teve um impacto profundo na indústria da música e na forma como consumimos música gerando um declínio das vendas de CDs: Os consumidores passaram a baixar e compartilhar músicas digitalmente. O MP3 impulsionou o crescimento da música digital, com o surgimento de serviços de streaming online como Spotify, Apple Music e Deezer. O MP3 abriu caminho para novos modelos de negócios na indústria da música, como a venda de músicas online e assinaturas de serviços de streaming.

Apesar de seu sucesso, o MP3 também enfrentou alguns desafios: Em taxas de bits mais baixas (como 64 kbps), a qualidade de som do MP3 pode ser perceptívelmente inferior à do áudio não compactado. Propriedades Intelectuais: Questões de direitos autorais e pirataria surgiram com a proliferação de arquivos MP3 compartilhados online. Formatos Alternativos: Novos formatos de áudio digital, como FLAC e AAC, oferecem maior qualidade de som ou taxas de compressão mais eficientes, mas ainda não alcançaram o mesmo nível de popularidade do MP3.

Outros padrões: MID (notação musical sintética), WAV (nenhuma perda de qualidade, logo arquivos grandes), além do formato OGG, da plataforma APPLE.

Vídeo No vídeo, estamos diante de necessidade maior de compressão. Se uma imagem estática já é grande, imagine precisar de 30 imagens estáticas por segundo. Agora, a localidade de referência ganha um novo viés, a l.r. no tempo. Ou seja, se imaginarmos um vídeo como um array 3-d (dimensões tempo, altura e largura), podemos comprimir em 2 dessas dimensões.

Aqui a história começa com os formatos analógicos (VHS, Betamax), segue pelo CD-ROM e VCD, e chega ao DVD. Depois vem os formatos AVC (padrão H.264), depois HEVC (H.265), VP9 (formato livre e aberto da Google) e AV1 (formato livre e aberto, melhor que o H.265). Diferentemente da imagem (e semelhantemente ao som), aqui existe muita preocupação com a latência (atraso no envio dos pacotes) que impacta diretamente a sincronia. O padrão H.264 utiliza timestamps e buffers para sincronizar áudio e vídeo. Em caso de perda de pacotes, mecanismos de recuperação de erros como o Concealment Motion Vector (CMV) podem ser utilizados para minimizar o impacto na sincronia. O CMV divide o vídeo em macroblocos. Quando há perda de um pacote contendo os vetores de movimento de um determinado macrobloco, o CMV busca os macroblocos vizinhos ($\uparrow, \downarrow, \leftarrow \rightarrow$) e ve se algum deles tem movimento similar ao esperado do macrobloco perdido. Daí ele "empresta" vetores e reconstitui o movimento ausente.

PDF O Portable Document Format (PDF), criado pela Adobe em 1993, se tornou um dos formatos de arquivo mais utilizados no mundo, presente em diversos setores e aplicações. Mais detalhes: vivx680a. Um documento PDF pode ser aberto e visualizado em qualquer dispositivo com um leitor de PDF instalado, independentemente do sistema operacional ou hardware utilizado. Isso garante que a formatação original do documento é sempre a mesma, independentemente da plataforma em que ele é aberto. Isso é crucial para documentos que exigem alta fidelidade visual, como apresentações, relatórios e publicações. O PDF oferece recursos de segurança robustos, como criptografia e assinaturas digitais, que protegem o conteúdo contra acessos não autorizados, modificações e falsificações. Isso torna o PDF ideal para documentos confidenciais e contratos legais. Até porque ele pode ser acompanhado de uma assinatura digital (*hash*). Suporta recursos de acessibilidade, como tags de estrutura e legendas para imagens, que facilitam o acesso ao conteúdo do documento por pessoas com deficiência visual ou outras necessidades especiais. Ao contrário de outros formatos de arquivo que podem ser corrompidos ou se tornar obsoletos com o tempo, o PDF é um formato durável e estável. Pense na economia de árvores: imagine se todos os PDFs existentes fossem para o papel. Ele é ideal para compartilhar documentos com outras pessoas, pois garante que a formatação original seja preservada e o conteúdo seja visualizado corretamente em diferentes dispositivos. O PDF é uma ótima opção para armazenar documentos importantes de forma segura e durável, pois protege o conteúdo contra modificações e garante que ele possa ser acessado no futuro. Revistas, livros, relatórios e outros materiais de publicação podem ser distribuídos no formato PDF, garantindo uma experiência de leitura consistente em diferentes dispositivos. **Outros:** Deixo de tratar aqui, por absoluta falta de espaço, outros BLOBs como exames médicos, arquivos de telescópios (vivx760), ...

Olhando dentro de um BLOB

Agora e hora de examinar com um "microscópio" as partes internas de um desses arquivos. O escolhido pela importância é o arquivo PDF. Pode parecer pouco, mas esta ideia simples, já salvou centenas de milhares (milhões ?) de árvores. Já são muitos os sistemas de informação que não geram mais saídas em papel, limitando-se a criar arquivos PDF.

Um arquivo PDF descreve como uma página ficará depois de impressa e ele é melhor do que imagem dessa página pelas seguintes razões

- ⋆ A imagem como tal sempre é muito maior do que sua descrição. A razão para isso é a evidente redundância de qualquer imagem.

- ⋆ A imagem como tal é fixa e só pode ser rotacionada, redimensionada ou de alguma maneira modificada à custa de processamento e perda de qualidade.
- ⋆ Sobre uma imagem não há como manipular o conteúdo, por exemplo conduzindo uma busca textual (^F alguma coisa ou então ^C e ^V).
- ⋆ Alterar uma única palavra sobre uma imagem exige habilidades de pintura. Sobre a descrição de uma página é tarefa simples: basta trocar uma palavra pela outra.
- ⋆ Imagine uma imagem enviada para um monitor (72 DPP) ou a mesma imagem enviada para uma impressora (300 DPP). Não podem ser a mesma imagem.

A versão 1.0 do padrão PDF nasceu em 1993 quando a Adobe lançou 2 programas: o Distiller (para gerar PDF) e o Reader (para ler): ambos pagos. A coisa começou a mudar quando o departamento de rendas americano comprou uma licença que permitia aos contribuintes baixar o Reader. Rapidamente a Adobe deu-se conta de que seria melhor para ela distribuir gratuitamente o Reader. Nos dez anos seguintes o PDF acabou se tornando o padrão de fato de descrição de imagens.

Apresenta as seguintes vantagens:

- ⋆ acesso aleatório: qualquer página pode ser montada independente das demais.
- ⋆ linearização: todos os elementos necessários para montar uma página estão juntos.
- ⋆ atualização incremental: mudanças ficam registradas no fim do arquivo. Vantagem: opção de desfazer ilimitada e rápida atualização. Desvantagem: o arquivo não pára de crescer.
- ⋆ fontes incorporadas: o destino sempre será montado corretamente. A fonte é desbastada de tudo que não é necessário.
- ⋆ padronização ISO: em 2008 a ISO abraçou o padrão Adobe PDF 1.7, o que o transformou em padrão aberto.
- ⋆ compatibilidade para trás e para a frente: Qualquer leitor lê as versões antigas e deve ler as futuras, já que todos os leitores ignoram o que não conhecem.
- ⋆ integrado a todos os principais browsers do protocolo `HTTP`.

Um arquivo PDF

Cabeçalho Formado por 2 linhas: a primeira identifica o arquivo como um PDF e informa a versão PDF em que ele foi gerado. A segunda linha inclui caracteres que não podem ser impressos (para avisar eventuais leitores deste fato). Exemplo:

```
%PDF-1.0
%âã
```

Corpo Um conjunto de nodos de um grafo, contendo as páginas, conteúdo gráfico, textos, sempre na forma de objetos. Cada objeto começa com um número de objeto (iniciando em 1), um número de geração (sempre 0) e a palavra chave `obj`. O objeto termina pela palavra chave `endobj`. Exemplo:

```
1 0 obj     -- descrevendo o objeto 1
<<          -- começa um dicionário
 /Kids [2 0 R]    -- /Kids tem o valor 2 0 R
 /Count 1         -- só um
 /Type /Pages     -- do tipo : página
>>               -- fim do dicionário
endobj            -- fim do objeto
```

Tabela de Referência Cruzada Lista o deslocamento em bytes de cada objeto no corpo do arquivo. Exemplo

```
0 6          -- seis entradas na tabela
0000000000  65535  f   -- entrada especial
0000000015  00000  n   -- obj1 no desloc 15
0000000074  00000  n   -- obj2 no desloc 74
...
```

Rodapé Informações e metadados para aumentar a performance de acesso. A primeira entrada é `trailer`. Depois o dicionário de trailer que deve apresentar ao menos a entrada `\Size` dizendo quantos itens há na tabela de referência cruzada e `\Root` dizendo qual objeto é o catálogo do documento. Depois vem uma linha que só contém `startxref` seguida por uma única linha indicando o deslocamento em bytes utilizado no inicio da tabela de referência cruzada do arquivo. Termina tudo a linha `%%EOF`. Exemplo:

```
trailer   -- palavra chave
<<   -- começa um dicionário
/Root 5 0 R   -- o catálogo é o objeto 5
/Size 6       -- tem o tamanho 6
>>   -- fim do dicionário
startxref
459  -- deslocamento da tabela de ref cruzada
%%EOF   -- fim de arquivo
```

Componentes

Uma descrição PDF suporta 5 componentes básicos: i. Números inteiros e reais (mas não exponenciais); ii. strings que são escritos entre parênteses; iii.Nomes, que sempre começam por uma barra normal; iv.Os valores booleanos `true` e `false` e v. O componente nulo (`null`). Aceita também 3 componentes compostos: i. Arrays, que são uma coleção ordenada de outros componentes, escritos entre colchetes; ii.Dicionários que são uma lista não ordenada de duplas: nome e componente. Começam por `<<` e terminam por `>>`. Finalmente iii. Fluxos, que armazenam dados binários, sempre acompanhados de um dicionário que descreve seus atributos, como comprimento e parâmetros de compactação, por exemplo.

☞ Para você fazer

As coisas em um arquivo PDF são separadas por um divisor de linhas. Pode ser o caracter $X'10'$, ou o $X'13'$ ou uma combinação de ambos. Ao examinar um PDF em hexadecimal, acostume-se a separar as linhas marcando este caractere. Isto posto, os componentes principais do arquivo PDF são:

- Header: identificação PDF e uma coleção de letras acentuadas
- Corpo: objetos numerados a partir de 1 sequencialmente, e identificados pela palavra `obj` no início e `endobj` no final. Esses objetos podem ser textos, desenhos, fontes, imagens, metadados, links, e um monte de coisas mais.
- Tabela de referência: uma lista de onde (no arquivo) está cada objeto. Esta tabela permite o acesso direto a cada um e a todos os objetos.
- Rodapé: Diversas informações que permitem o acesso rápido ao arquivo.

Nos interessa, neste exercício a informação `startxref` que é uma das últimas do arquivo. Você deve abrir o arquivo

```
arq02.pdf
```

publicado no lugar usual, com um editor de hexadecimal e deve procurar o startxref no final do arquivo. Na linha seguinte, existe um endereço que o PDF coloca em decimal. Você precisa transformar este número em hexadecimal e localizar no arquivo (mais acima) este endereço.

Nele, deve aparecer a palavra `xref` e na linha seguinte um zero e um número. Na linha seguinte, uma entrada (que todo arquivo tem) com muitos zeros, o número 65535 e uma letra. Na linha seguinte, **um número** seguido de zeros e uma letra. Trata-se do endereço do primeiro objeto do arquivo.

Responda aqui:

endereço da xref em hexadecimal	endereço do primeiro objeto em decimal

CEP - UP - UTFPR - PUCPr - UFPR –
17/06/2024 - 17:11:23.4
Prof Dr P Kantek (pkantek@gmail.com)
Objetos binários grandes vivxq10a, V: 1.01 76551
GABRIELA BERNARDO PASQUALOTTO
24CC2301 - 7 entregar ate 4/julho _____ /
/ _____

Objetos binários grandes - BLOBs

O objetivo desta atividade e conhecer e manusear os chamados objetos binários grandes: imagens, vídeos, documentos, músicas, etc. Tais objetos quando estão na memória do computador são conhecidos como BLOBs e quando vão para uma memória não volátil (pendrives, discos, SSDs, nuvem, ...) passam a ser arquivos.

Resultados da digitalização crescente da nossa sociedade, precisam ser conhecidos para deles se extraírem todos os benefícios que eles apresentam.

Eis alguns raciocínios: a documentação em papel de um avião boeing 747 ocuparia 5 aviões 747 para ser transportada. Alguém ainda compra enciclopédias em papel ? Lembram de quando as fotografias eram analógicas ? Um filme Ektachrome (36 fotos) custava bem caro. Vou chutar uns 150 reais, o que daria quase 5 reais/foto, só o filme, sem contar a revelação e as ampliações. Quando o I.R. era em papel, a receita precisava montar uma operação de guerra, envolvendo todos os bancos, apenas para receber as declarações, e isso que era mais ou menos 1/3 do número de contribuintes atual. E as músicas: durante muitos anos, no dia do meu pagamento eu ia a uma loja de discos (?) e comprava 1 único LP, ($\pm$ 30 minutos de música) era o que o meu dinheiro alcançava. Aliás, falando de música, vale comentar a ascensão e queda do formato CD. Ele passou de maravilha tecnológica em meados dos 80 a algo obsoleto em meados dos 2010. Do céu ao inferno em 30 anos. E, se perder em uma cidade estranha, tendo apenas o guia 4 Rodas para ajudar ? Uma sensação indescritível.

Vamos estudar mais de perto alguns BLOBs, lembrando que qualquer um pode criar e chamar de seu, um BLOB. Pela sua definição (grande bloco de dados binários com uma lógica interna de organização e acesso) vê-se que isto é perfeitamente possível. Assim, vamos nos limitar a BLOBs padrão como formatos e conceitos conhecidos e publicados.

Imagem É difícil contar esta história. Ela começa no daguerrótipo de meados do século 19, através da fixação de imagens em placas de vidro. Segue pelas imagens de TV que eram capturadas e transmitidas eletronicamente. Seu registro foi feito em gravações magnéticas analógicas. Uma nova necessidade surgiu na obtenção de fotos na Lua e além, quando não havia como retornar filmes. A propósito, um pouco antes disto, a vigilância eletrônica na Guerra Fria era muito custosa: satélites fotografavam o inimigo e precisavam ejetar os filmes (que acabavam rápido) ao sobrevoar território amigo. A primeira foto da lua, transmitida eletronicamente foi feita pela espaçonave Ranger em 1964, 17 minutos antes dela se espatifar na superfície lunar. Quando a Internet começou (por volta de 1990) a necessidade de manusear imagens sofreu uma mudança: antes o gerador e o visualizador eram de mesma origem, pelo que não havia necessidade de padronização. Com a Internet essa história mudou. Uma das primeiras respostas a esta demanda foi o formato BMP (Bitmap Image File). Embora originalmente proposto pela Microsoft e IBM, teve seu padrão publicado e virou de fato, um formato livre. Sua origem é a bio-física dos olhos humanos (3 canais separados para Red, Green e Blue), além de um truque bem legal chamado mapeamento pelo qual se negocia tamanho do arquivo VERSUS menor quantidade de cores. A principal inovação do BMP foi criar, publicar e obedecer a um padrão. Parece pouco, mas não é. A chave da popularização e do sucesso sempre é um PADRãO. Anote isso na sua cabeça e nunca mais esqueça. O padrão pode ser ruim ou bom (neste caso não é muito), mas é um padrão. A imagem é uma matriz numérica (outra inovação entusiasmadora), logo sujeita a sofrer a ação maravilhosa de qualquer matemática.

Padrão BMP O arquivo começa com um bloco de controle de 54 bytes, que contém o identificador 'BM', o tamanho do objeto (arquivo), onde a imagem começa, L=largura e A=altura da imagem, profundidade). Depois, dependendo da profundidade, pode haver uma tabela de cores. Finalmente a imagem na forma de $L \times A$ se for uma imagem mapeada ou 3 tabelas de $L \times A$, uma para cada cor se for uma imagem true-color. Veja na imagem a seguir estes conceitos

```
Exemplo:
        0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
        -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
0000    42 4D 62 05 00 00 00 00 00 00 36 04 00 00 28 00 BM.....6...(
0010    00 00 12 00 00 00 0F 00 00 00 01 00 08 00 00 00 ................
0020    00 00 60 01 00 00 00 00 00 00 00 00 00 00 00 01 ................
0030    00 00 00 00 00 00 00 00 FF FF 00 80 80 80 00 FF FF ............
0040    FF 00 FF 00 FF 00 00 00 FF 00 00 00 00 00 00 FF ................
0050    00 00 FF FF 00 00 FF 00 00 00 09 09 09 00 0A 0A ................
0060    0A 00 0B 0B 0B 00 0C 0C 0C 00 0D 0D 0D 00 0E 0E ................
0070    0E 00 0F 0F 0F 00 10 10 10 00 11 11 11 00 12 12 ................
...
0410    F6 00 F7 F7 F7 00 F8 F8 F8 00 F9 F9 F9 00 FA FA ................
0420    FA 00 FB FB FB 00 FC FC FC 00 FD FD FD 00 FE FE ................
0430    FE 00 FF FF FF 00 01 01 03 01 01 01 08 01 01 01 ................
0440    01 01 02 01 01 06 01 01 00 00 04 04 03 04 04 04 ................
0450    04 04 04 04 04 04 04 04 04 04 04 04 00 00 01 01 ................
```

Os problemas do BMP incluem: nenhuma compressão, o que graças ao fenômeno da *localidade de referência espacial* é uma tragédia. Também não prevê nenhum tipo de animação, o que faz falta na Internet. Não sabia (hoje sabe) lidar com transparência.

Um concorrente importante do BMP (lá nos primórdios) foi o formato GIF que admite compressão (usando o belo algoritmo LZW) e também admite uma forma rudimentar de animação. O problema do GIF é que ele foi e é um formato proprietário. Isto deu origem ao formato PNG que é praticamente o mesmo, mas livre.

A compressão poderia ser muito melhor, se se admitisse algum nivel de degradação na reconstituição da imagem após sua transformação em BLOB. Isto foi conseguido no padrão JPEG, que ao admitir degradações (sob controle), gera compressões muito maiores do que GIF ou similares. Outra vantagem é a possibilidade de criação em etapas das imagens (coisa que o BMP nunca ofereceu). O padrão foi formalizado pela ISO em 1991, mas na época disparou uma briga por patentes e direitos. Nos EUA, as últimas patentes expiraram em 2004. A compressão aqui é pelo uso da Transformada Discreta do Coseno. A sua operação está além do nosso interesse (querendo, olhe VIVX690c ou 690p), mas em resumo: a imagem é quebrada em blocos de 8×8; esta matriz é multiplicada pela matriz DCT, o que passa do domínio do pixel para o domínio da frequência (Transformada de Fourier); os coeficientes são quantizados; os coeficientes são comprimidos. Depois, se faz o inverso e *voilà*.

Som O MP3, sigla para MPEG-1 Audio Layer III, é um formato de arquivo de áudio digital que revolucionou a forma como consumimos e compartilhamos música.

Desenvolvido no início da década de 1990 pelo Moving Picture Experts Group (MPEG), o MP3 se tornou o formato dominante para música digital, utilizado em players de música portáteis, computadores, smartphones e serviços de streaming online. O MP3 utiliza uma técnica de compressão com perda chamada percepção auditiva psicoacústica para reduzir o tamanho dos arquivos de áudio. Essa técnica remove informações da faixa de áudio que o ouvido humano provavelmente não consegue perceber, resultando em arquivos menores sem comprometer significativamente a qualidade da música para a maioria dos ouvintes. Apesar da compressão, o MP3 oferece uma qualidade de som bastante boa, especialmente em taxas de bits mais altas (como 128 kbps ou superior). Tamanho de Arquivo Reduzido: Os arquivos MP3 são significativamente menores do que os arquivos de áudio não compactados, como CDs, permitindo fácil armazenamento e compartilhamento de música digital.

O MP3 teve um impacto profundo na indústria da música e na forma como consumimos música gerando um declínio das vendas de CDs: Os consumidores passaram a baixar e compartilhar músicas digitalmente. O MP3 impulsionou o crescimento da música digital, com o surgimento de serviços de streaming online como Spotify, Apple Music e Deezer. O MP3 abriu caminho para novos modelos de negócios na indústria da música, como a venda de músicas online e assinaturas de serviços de streaming.

Apesar de seu sucesso, o MP3 também enfrentou alguns desafios: Em taxas de bits mais baixas (como 64 kbps), a qualidade de som do MP3 pode ser perceptívelmente inferior à do áudio não compactado. Propriedades Intelectuais: Questões de direitos autorais e pirataria surgiram com a proliferação de arquivos MP3 compartilhados online. Formatos Alternativos: Novos formatos de áudio digital, como FLAC e AAC, oferecem maior qualidade de som ou taxas de compressão mais eficientes, mas ainda não alcançaram o mesmo nível de popularidade do MP3.

Outros padrões: MID (notação musical sintética), WAV (nenhuma perda de qualidade, logo arquivos grandes), além do formato OGG, da plataforma APPLE.

Vídeo No vídeo, estamos diante de necessidade maior de compressão. Se uma imagem estática já é grande, imagine precisar de 30 imagens estáticas por segundo. Agora, a localidade de referência ganha um novo viés, a l.r. no tempo. Ou seja, se imaginarmos um vídeo como um array 3-d (dimensões tempo, altura e largura), podemos comprimir em 2 dessas dimensões.

Aqui a história começa com os formatos analógicos (VHS, Betamax), segue pelo CD-ROM e VCD, e chega ao DVD. Depois vem os formatos AVC (padrão H.264), depois HEVC (H.265), VP9 (formato livre e aberto da Google) e AV1 (formato livre e aberto, melhor que o H.265). Diferentemente da imagem (e semelhantemente ao som), aqui existe muita preocupação com a latência (atraso no envio dos pacotes) que impacta diretamente a sincronia. O padrão H.264 utiliza timestamps e buffers para sincronizar áudio e vídeo. Em caso de perda de pacotes, mecanismos de recuperação de erros como o Concealment Motion Vector (CMV) podem ser utilizados para minimizar o impacto na sincronia. O CMV divide o vídeo em macroblocos. Quando há perda de um pacote contendo os vetores de movimento de um determinado macrobloco, o CMV busca os macroblocos vizinhos ($\uparrow, \downarrow, \leftarrow \rightarrow$) e ve se algum deles tem movimento similar ao esperado do macrobloco perdido. Daí ele "empresta" vetores e reconstitui o movimento ausente.

PDF O Portable Document Format (PDF), criado pela Adobe em 1993, se tornou um dos formatos de arquivo mais utilizados no mundo, presente em diversos setores e aplicações. Mais detalhes: vivx680a. Um documento PDF pode ser aberto e visualizado em qualquer dispositivo com um leitor de PDF instalado, independentemente do sistema operacional ou hardware utilizado. Isso garante que a formatação original do documento é sempre a mesma, independentemente da plataforma em que ele é aberto. Isso é crucial para documentos que exigem alta fidelidade visual, como apresentações, relatórios e publicações. O PDF oferece recursos de segurança robustos, como criptografia e assinaturas digitais, que protegem o conteúdo contra acessos não autorizados, modificações e falsificações. Isso torna o PDF ideal para documentos confidenciais e contratos legais. Até porque ele pode ser acompanhado de uma assinatura digital (*hash*). Suporta recursos de acessibilidade, como tags de estrutura e legendas para imagens, que facilitam o acesso ao conteúdo do documento por pessoas com deficiência visual ou outras necessidades especiais. Ao contrário de outros formatos de arquivo que podem ser corrompidos ou se tornar obsoletos com o tempo, o PDF é um formato durável e estável. Pense na economia de árvores: imagine se todos os PDFs existentes fossem para o papel. Ele é ideal para compartilhar documentos com outras pessoas, pois garante que a formatação original seja preservada e o conteúdo seja visualizado corretamente em diferentes dispositivos. O PDF é uma ótima opção para armazenar documentos importantes de forma segura e durável, pois protege o conteúdo contra modificações e garante que ele possa ser acessado no futuro. Revistas, livros, relatórios e outros materiais de publicação podem ser distribuídos no formato PDF, garantindo uma experiência de leitura consistente em diferentes dispositivos. **Outros:** Deixo de tratar aqui, por absoluta falta de espaço, outros BLOBs como exames médicos, arquivos de telescópios (vivx760), ...

Olhando dentro de um BLOB

Agora e hora de examinar com um "microscópio" as partes internas de um desses arquivos. O escolhido pela importância é o arquivo PDF. Pode parecer pouco, mas esta ideia simples, já salvou centenas de milhares (milhões ?) de árvores. Já são muitos os sistemas de informação que não geram mais saídas em papel, limitando-se a criar arquivos PDF.

Um arquivo PDF descreve como uma página ficará depois de impressa e ele é melhor do que imagem dessa página pelas seguintes razões

⋆ A imagem como tal sempre é muito maior do que

sua descrição. A razão para isso é a evidente redundância de qualquer imagem.

⋆ A imagem como tal é fixa e só pode ser rotacionada, redimensionada ou de alguma maneira modificada à custa de processamento e perda de qualidade.

⋆ Sobre uma imagem não há como manipular o conteúdo, por exemplo conduzindo uma busca textual (^F alguma coisa ou então ^C e ^V).

⋆ Alterar uma única palavra sobre uma imagem exige habilidades de pintura. Sobre a descrição de uma página é tarefa simples: basta trocar uma palavra pela outra.

⋆ Imagine uma imagem enviada para um monitor (72 DPP) ou a mesma imagem enviada para uma impressora (300 DPP). Não podem ser a mesma imagem.

A versão 1.0 do padrão PDF nasceu em 1993 quando a Adobe lançou 2 programas: o Distiller (para gerar PDF) e o Reader (para ler): ambos pagos. A coisa começou a mudar quando o departamento de rendas americano comprou uma licença que permitia aos contribuintes baixar o Reader. Rapidamente a Adobe deu-se conta de que seria melhor para ela distribuir gratuitamente o Reader. Nos dez anos seguintes o PDF acabou se tornando o padrão de fato de descrição de imagens.

Apresenta as seguintes vantagens:

⋆ acesso aleatório: qualquer página pode ser montada independente das demais.

⋆ linearização: todos os elementos necessários para montar uma página estão juntos.

⋆ atualização incremental: mudanças ficam registradas no fim do arquivo. Vantagem: opção de desfazer ilimitada e rápida atualização. Desvantagem: o arquivo não pára de crescer.

⋆ fontes incorporadas: o destino sempre será montado corretamente. A fonte é desbastada de tudo que não é necessário.

⋆ padronização ISO: em 2008 a ISO abraçou o padrão Adobe PDF 1.7, o que o transformou em padrão aberto.

⋆ compatibilidade para trás e para a frente: Qualquer leitor lê as versões antigas e deve ler as futuras, já que todos os leitores ignoram o que não conhecem.

⋆ integrado a todos os principais browsers do protocolo `HTTP`.

Um arquivo PDF

Cabeçalho Formado por 2 linhas: a primeira identifica o arquivo como um PDF e informa a versão PDF em que ele foi gerado. A segunda linha inclui caracteres que não podem ser impressos (para avisar eventuais leitores deste fato). Exemplo:

```
%PDF-1.0
%âã
```

Corpo Um conjunto de nodos de um grafo, contendo as páginas, conteúdo gráfico, textos, sempre na forma de objetos. Cada objeto começa com um número de objeto (iniciando em 1), um número de geração (sempre 0) e a palavra chave `obj`. O objeto termina pela palavra chave `endobj`. Exemplo:

```
1 0 obj     -- descrevendo o objeto 1
<<          -- começa um dicionário
 /Kids [2 0 R]     -- /Kids tem o valor 2 0 R
 /Count 1          -- só um
 /Type /Pages      -- do tipo : página
>>                 -- fim do dicionário
endobj             -- fim do objeto
```

Tabela de Referência Cruzada Lista o deslocamento em bytes de cada objeto no corpo do arquivo. Exemplo

```
0 6           -- seis entradas na tabela
0000000000  65535  f   -- entrada especial
0000000015  00000  n   -- obj1 no desloc 15
0000000074  00000  n   -- obj2 no desloc 74
...
```

Rodapé Informações e metadados para aumentar a performance de acesso. A primeira entrada é `trailer`. Depois o dicionário de trailer que deve apresentar ao menos a entrada `\Size` dizendo quantos itens há na tabela de referência cruzada e `\Root` dizendo qual objeto é o catálogo do documento. Depois vem uma linha que só contém `startxref` seguida por uma única linha indicando o deslocamento em bytes utilizado no inicio da tabela de referência cruzada do arquivo. Termina tudo a linha `%%EOF`. Exemplo:

```
trailer   -- palavra chave
<<   -- começa um dicionário
/Root 5 0 R   -- o catálogo é o objeto 5
/Size 6       -- tem o tamanho 6
>>   -- fim do dicionário
startxref
459  -- deslocamento da tabela de ref cruzada
%%EOF   -- fim de arquivo
```

Componentes

Uma descrição PDF suporta 5 componentes básicos: i. Números inteiros e reais (mas não exponenciais); ii. strings que são escritos entre parênteses; iii.Nomes, que sempre começam por uma barra normal; iv.Os valores booleanos `true` e `false` e v. O componente nulo (`null`). Aceita também 3 componentes compostos: i. Arrays, que são uma coleção ordenada de outros componentes, escritos entre colchetes; ii.Dicionários que são uma lista não ordenada de duplas: nome e componente. Começam por `<<` e terminam por `>>`. Finalmente iii. Fluxos, que armazenam dados binários, sempre acompanhados de um dicionário que descreve seus atributos, como comprimento e parâmetros de compactação, por exemplo.

☞ Para você fazer

As coisas em um arquivo PDF são separadas por um divisor de linhas. Pode ser o caracter $X'10'$, ou o $X'13'$ ou uma combinação de ambos. Ao examinar um PDF em hexadecimal, acostume-se a separar as linhas marcando este caractere. Isto posto, os componentes principais do arquivo PDF são:

- Header: identificação PDF e uma coleção de letras acentuadas
- Corpo: objetos numerados a partir de 1 sequencialmente, e identificados pela palavra `obj` no início e `endobj` no final. Esses objetos podem ser textos, desenhos, fontes, imagens, metadados, links, e um monte de coisas mais.
- Tabela de referência: uma lista de onde (no arquivo) está cada objeto. Esta tabela permite o acesso direto a cada um e a todos os objetos.
- Rodapé: Diversas informações que permitem o acesso rápido ao arquivo.

Nos interessa, neste exercício a informação `startxref` que é uma das últimas do arquivo. Você deve abrir o arquivo

```
arq09.pdf
```

publicado no lugar usual, com um editor de hexadecimal e deve procurar o startxref no final do arquivo. Na linha seguinte, existe um endereço que o PDF coloca em decimal. Você precisa transformar este número em hexadecimal e localizar no arquivo (mais acima) este endereço.

Nele, deve aparecer a palavra `xref` e na linha seguinte um zero e um número. Na linha seguinte, uma entrada (que todo arquivo tem) com muitos zeros, o número 65535 e uma letra. Na linha seguinte, **um número** seguido de zeros e uma letra. Trata-se do endereço do primeiro objeto do arquivo.

Responda aqui:

endereço da xref em hexadecimal	endereço do primeiro objeto em decimal

CEP - UP - UTFPR - PUCPr - UFPR –
17/06/2024 - 17:11:23.4
Prof Dr P Kantek (pkantek@gmail.com)
Objetos binários grandes vivxq10a, V: 1.01 76568
GUILHERME VITOR DA SILVA 24CC2301 - 8
entregar ate 4/julho _____ / _____ / _____

Objetos binários grandes - BLOBs

O objetivo desta atividade e conhecer e manusear os chamados objetos binários grandes: imagens, vídeos, documentos, músicas, etc. Tais objetos quando estão na memória do computador são conhecidos como BLOBs e quando vão para uma memória não volátil (pendrives, discos, SSDs, nuvem, ...) passam a ser arquivos.

Resultados da digitalização crescente da nossa sociedade, precisam ser conhecidos para deles se extraírem todos os benefícios que eles apresentam.

Eis alguns raciocínios: a documentação em papel de um avião boeing 747 ocuparia 5 aviões 747 para ser transportada. Alguém ainda compra enciclopédias em papel ? Lembram de quando as fotografias eram analógicas ? Um filme Ektachrome (36 fotos) custava bem caro. Vou chutar uns 150 reais, o que daria quase 5 reais/foto, só o filme, sem contar a revelação e as ampliações. Quando o I.R. era em papel, a receita precisava montar uma operação de guerra, envolvendo todos os bancos, apenas para receber as declarações, e isso que era mais ou menos 1/3 do número de contribuintes atual. E as músicas: durante muitos anos, no dia do meu pagamento eu ia a uma loja de discos (?) e comprava 1 único LP, ($\pm$ 30 minutos de música) era o que o meu dinheiro alcançava. Aliás, falando de música, vale comentar a ascensão e queda do formato CD. Ele passou de maravilha tecnológica em meados dos 80 a algo obsoleto em meados dos 2010. Do céu ao inferno em 30 anos. E, se perder em uma cidade estranha, tendo apenas o guia 4 Rodas para ajudar ? Uma sensação indescritível.

Vamos estudar mais de perto alguns BLOBs, lembrando que qualquer um pode criar e chamar de seu, um BLOB. Pela sua definição (grande bloco de dados binários com uma lógica interna de organização e acesso) vê-se que isto é perfeitamente possível. Assim, vamos nos limitar a BLOBs padrão como formatos e conceitos conhecidos e publicados.

Imagem É difícil contar esta história. Ela começa no daguerrótipo de meados do século 19, através da fixação de imagens em placas de vidro. Segue pelas imagens de TV que eram capturadas e transmitidas eletronicamente. Seu registro foi feito em gravações magnéticas analógicas. Uma nova necessidade surgiu na obtenção de fotos na Lua e além, quando não havia como retornar filmes. A propósito, um pouco antes disto, a vigilância eletrônica na Guerra Fria era muito custosa: satélites fotografavam o inimigo e precisavam ejetar os filmes (que acabavam rápido) ao sobrevoar território amigo. A primeira foto da lua, transmitida eletronicamente foi feita pela espaçonave Ranger em 1964, 17 minutos antes dela se espatifar na superfície lunar. Quando a Internet começou (por volta de 1990) a necessidade de manusear imagens sofreu uma mudança: antes o gerador e o visualizador eram de mesma origem, pelo que não havia necessidade de padronização. Com a Internet essa história mudou. Uma das primeiras respostas a esta demanda foi o formato BMP (Bitmap Image File). Embora originalmente proposto pela Microsoft e IBM, teve seu padrão publicado e virou de fato, um formato livre. Sua origem é a bio-física dos olhos humanos (3 canais separados para Red, Green e Blue), além de um truque bem legal chamado mapeamento pelo qual se negocia tamanho do arquivo VERSUS menor quantidade de cores. A principal inovação do BMP foi criar, publicar e obedecer a um padrão. Parece pouco, mas não é. A chave da popularização e do sucesso sempre é um PADRãO. Anote isso na sua cabeça e nunca mais esqueça. O padrão pode ser ruim ou bom (neste caso não é muito), mas é um padrão. A imagem é uma matriz numérica (outra inovação entusiasmadora), logo sujeita a sofrer a ação maravilhosa de qualquer matemática.

Padrão BMP O arquivo começa com um bloco de controle de 54 bytes, que contém o identificador 'BM', o tamanho do objeto (arquivo), onde a imagem começa, L=largura e A=altura da imagem, profundidade). Depois, dependendo da profundidade, pode haver uma tabela de cores. Finalmente a imagem na forma de $L \times A$ se for uma imagem mapeada ou 3 tabelas de $L \times A$, uma para cada cor se for uma imagem true-color. Veja na imagem a seguir estes conceitos

Exemplo:

```
      0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
      -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
0000  42 4D 62 05 00 00 00 00 00 00 36 04 00 00 28 00 BM.....6...(
0010  00 00 12 00 00 00 0F 00 00 00 01 00 08 00 00 00 ................
0020  00 00 60 01 00 00 00 00 00 00 00 00 00 00 00 00 ..`.............
0030  00 00 00 00 00 00 00 00 FF FF 00 80 80 80 00 FF ................
0040  FF 00 FF 00 FF 00 00 00 00 FF 00 00 00 00 00 00 ................
0050  00 00 FF FF 00 00 FF 00 00 00 09 09 09 00 0A 0A ................
0060  0A 00 0B 0B 0B 00 0C 0C 0C 00 0D 0D 0D 00 0E 0E ................
0070  0E 00 0F 0F 0F 00 10 10 10 00 11 11 11 00 12 12 ................
...
0410  F6 00 F7 F7 F7 00 F8 F8 F8 00 F9 F9 F9 00 FA FA ................
0420  FA 00 FB FB FB 00 FC FC FC 00 FD FD FD 00 FE FE ................
0430  FE 00 FF FF FF 00 01 01 03 01 01 01 08 01 01 01 ................
0440  01 01 02 01 01 06 01 01 00 00 04 04 03 04 04 04 ................
0450  04 04 04 04 04 04 04 04 04 04 04 04 00 00 01 01 ................
```

Os problemas do BMP incluem: nenhuma compressão, o que graças ao fenômeno da *localidade de referência espacial* é uma tragédia. Também não prevê nenhum tipo de animação, o que faz falta na Internet. Não sabia (hoje sabe) lidar com transparência.

Um concorrente importante do BMP (lá nos primórdios) foi o formato GIF que admite compressão (usando o belo algoritmo LZW) e também admite uma forma rudimentar de animação. O problema do GIF é que ele foi e é um formato proprietário. Isto deu origem ao formato PNG que é praticamente o mesmo, mas livre.

A compressão poderia ser muito melhor, se se admitisse algum nivel de degradação na reconstituição da imagem após sua transformação em BLOB. Isto foi conseguido no padrão JPEG, que ao admitir degradações (sob controle), gera compressões muito maiores do que GIF ou similares. Outra vantagem é a possibilidade de criação em etapas das imagens (coisa que o BMP nunca ofereceu). O padrão foi formalizado pela ISO em 1991, mas na época disparou uma briga por patentes e direitos. Nos EUA, as últimas patentes expiraram em 2004. A compressão aqui é pelo uso da Transformada Discreta do Coseno. A sua operação está além do nosso interesse (querendo, olhe VIVX690c ou 690p), mas em resumo: a imagem é quebrada em blocos de 8×8; esta matriz é multiplicada pela matriz DCT, o que passa do domínio do pixel para o domínio da frequência (Transformada de Fourier); os coeficientes são quantizados; os coeficientes são comprimidos. Depois, se faz o inverso e *voilà*.

Som O MP3, sigla para MPEG-1 Audio Layer III, é um formato de arquivo de áudio digital que revolucionou a forma como consumimos e compartilhamos música.

Desenvolvido no início da década de 1990 pelo Moving Picture Experts Group (MPEG), o MP3 se tornou o formato dominante para música digital, utilizado em players de música portáteis, computadores, smartphones e serviços de streaming online. O MP3 utiliza uma técnica de compressão com perda chamada percepção auditiva psicoacústica para reduzir o tamanho dos arquivos de áudio. Essa técnica remove informações da faixa de áudio que o ouvido humano provavelmente não consegue perceber, resultando em arquivos menores sem comprometer significativamente a qualidade da música para a maioria dos ouvintes. Apesar da compressão, o MP3 oferece uma qualidade de som bastante boa, especialmente em taxas de bits mais altas (como 128 kbps ou superior). Tamanho de Arquivo Reduzido: Os arquivos MP3 são significativamente menores do que os arquivos de áudio não compactados, como CDs, permitindo fácil armazenamento e compartilhamento de música digital.

O MP3 teve um impacto profundo na indústria da música e na forma como consumimos música gerando um declínio das vendas de CDs: Os consumidores passaram a baixar e compartilhar músicas digitalmente. O MP3 impulsionou o crescimento da música digital, com o surgimento de serviços de streaming online como Spotify, Apple Music e Deezer. O MP3 abriu caminho para novos modelos de negócios na indústria da música, como a venda de músicas online e assinaturas de serviços de streaming.

Apesar de seu sucesso, o MP3 também enfrentou alguns desafios: Em taxas de bits mais baixas (como 64 kbps), a qualidade de som do MP3 pode ser perceptívelmente inferior à do áudio não compactado. Propriedades Intelectuais: Questões de direitos autorais e pirataria surgiram com a proliferação de arquivos MP3 compartilhados online. Formatos Alternativos: Novos formatos de áudio digital, como FLAC e AAC, oferecem maior qualidade de som ou taxas de compressão mais eficientes, mas ainda não alcançaram o mesmo nível de popularidade do MP3.

Outros padrões: MID (notação musical sintética), WAV (nenhuma perda de qualidade, logo arquivos grandes), além do formato OGG, da plataforma APPLE.

Vídeo No vídeo, estamos diante de necessidade maior de compressão. Se uma imagem estática já é grande, imagine precisar de 30 imagens estáticas por segundo. Agora, a localidade de referência ganha um novo viés, a l.r. no tempo. Ou seja, se imaginarmos um vídeo como um array 3-d (dimensões tempo, altura e largura), podemos comprimir em 2 dessas dimensões.

Aqui a história começa com os formatos analógicos (VHS, Betamax), segue pelo CD-ROM e VCD, e chega ao DVD. Depois vem os formatos AVC (padrão H.264), depois HEVC (H.265), VP9 (formato livre e aberto da Google) e AV1 (formato livre e aberto, melhor que o H.265). Diferentemente da imagem (e semelhantemente ao som), aqui existe muita preocupação com a latência (atraso no envio dos pacotes) que impacta diretamente a sincronia. O padrão H.264 utiliza timestamps e buffers para sincronizar áudio e vídeo. Em caso de perda de pacotes, mecanismos de recuperação de erros como o Concealment Motion Vector (CMV) podem ser utilizados para minimizar o impacto na sincronia. O CMV divide o vídeo em macroblocos. Quando há perda de um pacote contendo os vetores de movimento de um determinado macrobloco, o CMV busca os macroblocos vizinhos ($\uparrow, \downarrow, \leftarrow \rightarrow$) e ve se algum deles tem movimento similar ao esperado do macrobloco perdido. Daí ele "empresta" vetores e reconstitui o movimento ausente.

PDF O Portable Document Format (PDF), criado pela Adobe em 1993, se tornou um dos formatos de arquivo mais utilizados no mundo, presente em diversos setores e aplicações. Mais detalhes: vivx680a. Um documento PDF pode ser aberto e visualizado em qualquer dispositivo com um leitor de PDF instalado, independentemente do sistema operacional ou hardware utilizado. Isso garante que a formatação original do documento é sempre a mesma, independentemente da plataforma em que ele é aberto. Isso é crucial para documentos que exigem alta fidelidade visual, como apresentações, relatórios e publicações. O PDF oferece recursos de segurança robustos, como criptografia e assinaturas digitais, que protegem o conteúdo contra acessos não autorizados, modificações e falsificações. Isso torna o PDF ideal para documentos confidenciais e contratos legais. Até porque ele pode ser acompanhado de uma assinatura digital (*hash*). Suporta recursos de acessibilidade, como tags de estrutura e legendas para imagens, que facilitam o acesso ao conteúdo do documento por pessoas com deficiência visual ou outras necessidades especiais. Ao contrário de outros formatos de arquivo que podem ser corrompidos ou se tornar obsoletos com o tempo, o PDF é um formato durável e estável. Pense na economia de árvores: imagine se todos os PDFs existentes fossem para o papel. Ele é ideal para compartilhar documentos com outras pessoas, pois garante que a formatação original seja preservada e o conteúdo seja visualizado corretamente em diferentes dispositivos. O PDF é uma ótima opção para armazenar documentos importantes de forma segura e durável, pois protege o conteúdo contra modificações e garante que ele possa ser acessado no futuro. Revistas, livros, relatórios e outros materiais de publicação podem ser distribuídos no formato PDF, garantindo uma experiência de leitura consistente em diferentes dispositivos. **Outros:** Deixo de tratar aqui, por absoluta falta de espaço, outros BLOBs como exames médicos, arquivos de telescópios (vivx760), ...

Olhando dentro de um BLOB

Agora e hora de examinar com um "microscópio" as partes internas de um desses arquivos. O escolhido pela importância é o arquivo PDF. Pode parecer pouco, mas esta ideia simples, já salvou centenas de milhares (milhões ?) de árvores. Já são muitos os sistemas de informação que não geram mais saídas em papel, limitando-se a criar arquivos PDF.

Um arquivo PDF descreve como uma página ficará depois de impressa e ele é melhor do que imagem dessa página pelas seguintes razões

- ⋆ A imagem como tal sempre é muito maior do que sua descrição. A razão para isso é a evidente redundância de qualquer imagem.

⋆ A imagem como tal é fixa e só pode ser rotacionada, redimensionada ou de alguma maneira modificada à custa de processamento e perda de qualidade.

⋆ Sobre uma imagem não há como manipular o conteúdo, por exemplo conduzindo uma busca textual (^F alguma coisa ou então ^C e ^V).

⋆ Alterar uma única palavra sobre uma imagem exige habilidades de pintura. Sobre a descrição de uma página é tarefa simples: basta trocar uma palavra pela outra.

⋆ Imagine uma imagem enviada para um monitor (72 DPP) ou a mesma imagem enviada para uma impressora (300 DPP). Não podem ser a mesma imagem.

A versão 1.0 do padrão PDF nasceu em 1993 quando a Adobe lançou 2 programas: o Distiller (para gerar PDF) e o Reader (para ler): ambos pagos. A coisa começou a mudar quando o departamento de rendas americano comprou uma licença que permitia aos contribuintes baixar o Reader. Rapidamente a Adobe deu-se conta de que seria melhor para ela distribuir gratuitamente o Reader. Nos dez anos seguintes o PDF acabou se tornando o padrão de fato de descrição de imagens.

Apresenta as seguintes vantagens:

⋆ acesso aleatório: qualquer página pode ser montada independente das demais.

⋆ linearização: todos os elementos necessários para montar uma página estão juntos.

⋆ atualização incremental: mudanças ficam registradas no fim do arquivo. Vantagem: opção de desfazer ilimitada e rápida atualização. Desvantagem: o arquivo não pára de crescer.

⋆ fontes incorporadas: o destino sempre será montado corretamente. A fonte é desbastada de tudo que não é necessário.

⋆ padronização ISO: em 2008 a ISO abraçou o padrão Adobe PDF 1.7, o que o transformou em padrão aberto.

⋆ compatibilidade para trás e para a frente: Qualquer leitor lê as versões antigas e deve ler as futuras, já que todos os leitores ignoram o que não conhecem.

⋆ integrado a todos os principais browsers do protocolo `HTTP`.

Um arquivo PDF

Cabeçalho Formado por 2 linhas: a primeira identifica o arquivo como um PDF e informa a versão PDF em que ele foi gerado. A segunda linha inclui caracteres que não podem ser impressos (para avisar eventuais leitores deste fato). Exemplo:

```
%PDF-1.0
%âã
```

Corpo Um conjunto de nodos de um grafo, contendo as páginas, conteúdo gráfico, textos, sempre na forma de objetos. Cada objeto começa com um número de objeto (iniciando em 1), um número de geração (sempre 0) e a palavra chave `obj`. O objeto termina pela palavra chave `endobj`. Exemplo:

```
1 0 obj     -- descrevendo o objeto 1
<<          -- começa um dicionário
 /Kids [2 0 R]    -- /Kids tem o valor 2 0 R
 /Count 1         -- só um
 /Type /Pages     -- do tipo : página
>>               -- fim do dicionário
endobj            -- fim do objeto
```

Tabela de Referência Cruzada Lista o deslocamento em bytes de cada objeto no corpo do arquivo. Exemplo

```
0 6          -- seis entradas na tabela
0000000000  65535  f   -- entrada especial
0000000015  00000  n   -- obj1 no desloc 15
0000000074  00000  n   -- obj2 no desloc 74
...
```

Rodapé Informações e metadados para aumentar a performance de acesso. A primeira entrada é `trailer`. Depois o dicionário de trailer que deve apresentar ao menos a entrada `\Size` dizendo quantos itens há na tabela de referência cruzada e `\Root` dizendo qual objeto é o catálogo do documento. Depois vem uma linha que só contém `startxref` seguida por uma única linha indicando o deslocamento em bytes utilizado no inicio da tabela de referência cruzada do arquivo. Termina tudo a linha `%%EOF`. Exemplo:

```
trailer   -- palavra chave
<<   -- começa um dicionário
/Root 5 0 R   -- o catálogo é o objeto 5
/Size 6       -- tem o tamanho 6
>>   -- fim do dicionário
startxref
459  -- deslocamento da tabela de ref cruzada
%%EOF   -- fim de arquivo
```

Componentes

Uma descrição PDF suporta 5 componentes básicos: i. Números inteiros e reais (mas não exponenciais); ii. strings que são escritos entre parênteses; iii.Nomes, que sempre começam por uma barra normal; iv.Os valores booleanos `true` e `false` e v. O componente nulo (`null`). Aceita também 3 componentes compostos: i. Arrays, que são uma coleção ordenada de outros componentes, escritos entre colchetes; ii.Dicionários que são uma lista não ordenada de duplas: nome e componente. Começam por `<<` e terminam por `>>`. Finalmente iii. Fluxos, que armazenam dados binários, sempre acompanhados de um dicionário que descreve seus atributos, como comprimento e parâmetros de compactação, por exemplo.

☞ Para você fazer

As coisas em um arquivo PDF são separadas por um divisor de linhas. Pode ser o caracter $X'10'$, ou o $X'13'$ ou uma combinação de ambos. Ao examinar um PDF em hexadecimal, acostume-se a separar as linhas marcando este caractere. Isto posto, os componentes principais do arquivo PDF são:

- Header: identificação PDF e uma coleção de letras acentuadas
- Corpo: objetos numerados a partir de 1 sequencialmente, e identificados pela palavra `obj` no início e `endobj` no final. Esses objetos podem ser textos, desenhos, fontes, imagens, metadados, links, e um monte de coisas mais.
- Tabela de referência: uma lista de onde (no arquivo) está cada objeto. Esta tabela permite o acesso direto a cada um e a todos os objetos.
- Rodapé: Diversas informações que permitem o acesso rápido ao arquivo.

Nos interessa, neste exercício a informação `startxref` que é uma das últimas do arquivo. Você deve abrir o arquivo

```
arq08.pdf
```

publicado no lugar usual, com um editor de hexadecimal e deve procurar o startxref no final do arquivo. Na linha seguinte, existe um endereço que o PDF coloca em decimal. Você precisa transformar este número em hexadecimal e localizar no arquivo (mais acima) este endereço.

Nele, deve aparecer a palavra `xref` e na linha seguinte um zero e um número. Na linha seguinte, uma entrada (que todo arquivo tem) com muitos zeros, o número 65535 e uma letra. Na linha seguinte, **um número** seguido de zeros e uma letra. Trata-se do endereço do primeiro objeto do arquivo.

Responda aqui:

endereço da xref em hexadecimal	endereço do primeiro objeto em decimal

CEP - UP - UTFPR - PUCPr - UFPR –
17/06/2024 - 17:11:23.4
Prof Dr P Kantek (pkantek@gmail.com)
Objetos binários grandes vivxq10a, V: 1.01 76575
LAURA PACHECO DELATTRE 24CC2301 - 9
entregar ate 4/julho _____ / _____ / _____

Objetos binários grandes - BLOBs

O objetivo desta atividade e conhecer e manusear os chamados objetos binários grandes: imagens, vídeos, documentos, músicas, etc. Tais objetos quando estão na memória do computador são conhecidos como BLOBs e quando vão para uma memória não volátil (pendrives, discos, SSDs, nuvem, ...) passam a ser arquivos.

Resultados da digitalização crescente da nossa sociedade, precisam ser conhecidos para deles se extraírem todos os benefícios que eles apresentam.

Eis alguns raciocínios: a documentação em papel de um avião boeing 747 ocuparia 5 aviões 747 para ser transportada. Alguém ainda compra enciclopédias em papel ? Lembram de quando as fotografias eram analógicas ? Um filme Ektachrome (36 fotos) custava bem caro. Vou chutar uns 150 reais, o que daria quase 5 reais/foto, só o filme, sem contar a revelação e as ampliações. Quando o I.R. era em papel, a receita precisava montar uma operação de guerra, envolvendo todos os bancos, apenas para receber as declarações, e isso que era mais ou menos 1/3 do número de contribuintes atual. E as músicas: durante muitos anos, no dia do meu pagamento eu ia a uma loja de discos (?) e comprava 1 único LP, ($\pm$ 30 minutos de música) era o que o meu dinheiro alcançava. Aliás, falando de música, vale comentar a ascensão e queda do formato CD. Ele passou de maravilha tecnológica em meados dos 80 a algo obsoleto em meados dos 2010. Do céu ao inferno em 30 anos. E, se perder em uma cidade estranha, tendo apenas o guia 4 Rodas para ajudar ? Uma sensação indescritível.

Vamos estudar mais de perto alguns BLOBs, lembrando que qualquer um pode criar e chamar de seu, um BLOB. Pela sua definição (grande bloco de dados binários com uma lógica interna de organização e acesso) vê-se que isto é perfeitamente possível. Assim, vamos nos limitar a BLOBs padrão como formatos e conceitos conhecidos e publicados.

Imagem É difícil contar esta história. Ela começa no daguerrótipo de meados do século 19, através da fixação de imagens em placas de vidro. Segue pelas imagens de TV que eram capturadas e transmitidas eletronicamente. Seu registro foi feito em gravações magnéticas analógicas. Uma nova necessidade surgiu na obtenção de fotos na Lua e além, quando não havia como retornar filmes. A propósito, um pouco antes disto, a vigilância eletrônica na Guerra Fria era muito custosa: satélites fotografavam o inimigo e precisavam ejetar os filmes (que acabavam rápido) ao sobrevoar território amigo. A primeira foto da lua, transmitida eletronicamente foi feita pela espaçonave Ranger em 1964, 17 minutos antes dela se espatifar na superfície lunar. Quando a Internet começou (por volta de 1990) a necessidade de manusear imagens sofreu uma mudança: antes o gerador e o visualizador eram de mesma origem, pelo que não havia necessidade de padronização. Com a Internet essa história mudou. Uma das primeiras respostas a esta demanda foi o formato BMP (Bitmap Image File). Embora originalmente proposto pela Microsoft e IBM, teve seu padrão publicado e virou de fato, um formato livre. Sua origem é a bio-física dos olhos humanos (3 canais separados para Red, Green e Blue), além de um truque bem legal chamado mapeamento pelo qual se negocia tamanho do arquivo VERSUS menor quantidade de cores. A principal inovação do BMP foi criar, publicar e obedecer a um padrão. Parece pouco, mas não é. A chave da popularização e do sucesso sempre é um PADRãO. Anote isso na sua cabeça e nunca mais esqueça. O padrão pode ser ruim ou bom (neste caso não é muito), mas é um padrão. A imagem é uma matriz numérica (outra inovação entusiasmadora), logo sujeita a sofrer a ação maravilhosa de qualquer matemática.

Padrão BMP O arquivo começa com um bloco de controle de 54 bytes, que contém o identificador 'BM', o tamanho do objeto (arquivo), onde a imagem começa, L=largura e A=altura da imagem, profundidade). Depois, dependendo da profundidade, pode haver uma tabela de cores. Finalmente a imagem na forma de $L \times A$ se for uma imagem mapeada ou 3 tabelas de $L \times A$, uma para cada cor se for uma imagem true-color. Veja na imagem a seguir estes conceitos

```
Exemplo:
       0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
      -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
0000  42 4D 62 05 00 00 00 00 00 00 36 04 00 00 28 00 BM.....6...(
0010  00 00 12 00 00 00 0F 00 00 00 01 00 08 00 00 00 ................
0020  00 00 60 01 00 00 00 00 00 00 00 00 00 00 00 01 ..`.............
0030  00 00 00 00 00 00 00 00 FF FF 00 80 80 00 FF FF ................
0040  FF 00 FF 00 FF 00 00 00 00 FF 00 00 00 00 00 FF ................
0050  00 00 FF FF 00 00 FF 00 00 00 09 09 09 00 0A 0A ................
0060  0A 00 0B 0B 0B 00 0C 0C 0C 00 0D 0D 0D 00 0E 0E ................
0070  0E 00 0F 0F 0F 00 10 10 10 00 11 11 11 00 12 12 ................
...
0410  F6 00 F7 F7 F7 00 F8 F8 F8 00 F9 F9 F9 00 FA FA ................
0420  FA 00 FB FB FB 00 FC FC FC 00 FD FD FD 00 FE FE ................
0430  FE 00 FF FF FF 00 01 01 03 01 01 01 08 01 01 01 ................
0440  01 01 02 01 01 06 01 01 00 00 04 04 03 04 04 04 ................
0450  04 04 04 04 04 04 04 04 04 04 04 04 00 00 01 01 ................
```

Os problemas do BMP incluem: nenhuma compressão, o que graças ao fenômeno da *localidade de referência espacial* é uma tragédia. Também não prevê nenhum tipo de animação, o que faz falta na Internet. Não sabia (hoje sabe) lidar com transparência.

Um concorrente importante do BMP (lá nos primórdios) foi o formato GIF que admite compressão (usando o belo algoritmo LZW) e também admite uma forma rudimentar de animação. O problema do GIF é que ele foi e é um formato proprietário. Isto deu origem ao formato PNG que é praticamente o mesmo, mas livre.

A compressão poderia ser muito melhor, se se admitisse algum nivel de degradação na reconstituição da imagem após sua transformação em BLOB. Isto foi conseguido no padrão JPEG, que ao admitir degradações (sob controle), gera compressões muito maiores do que GIF ou similares. Outra vantagem é a possibilidade de criação em etapas das imagens (coisa que o BMP nunca ofereceu). O padrão foi formalizado pela ISO em 1991, mas na época disparou uma briga por patentes e direitos. Nos EUA, as últimas patentes expiraram em 2004. A compressão aqui é pelo uso da Transformada Discreta do Coseno. A sua operação está além do nosso interesse (querendo, olhe VIVX690c ou 690p), mas em resumo: a imagem é quebrada em blocos de 8×8; esta matriz é multiplicada pela matriz DCT, o que passa do domínio do pixel para o domínio da frequência (Transformada de Fourier); os coeficientes são quantizados; os coeficientes são comprimidos. Depois, se faz o inverso e *voilà*.

Som O MP3, sigla para MPEG-1 Audio Layer III, é um formato de arquivo de áudio digital que revolucionou a forma como consumimos e compartilhamos música.

Desenvolvido no início da década de 1990 pelo Moving Picture Experts Group (MPEG), o MP3 se tornou o formato dominante para música digital, utilizado em players de música portáteis, computadores, smartphones e serviços de streaming online. O MP3 utiliza uma técnica de compressão com perda chamada percepção auditiva psicoacústica para reduzir o tamanho dos arquivos de áudio. Essa técnica remove informações da faixa de áudio que o ouvido humano provavelmente não consegue perceber, resultando em arquivos menores sem comprometer significativamente a qualidade da música para a maioria dos ouvintes. Apesar da compressão, o MP3 oferece uma qualidade de som bastante boa, especialmente em taxas de bits mais altas (como 128 kbps ou superior). Tamanho de Arquivo Reduzido: Os arquivos MP3 são significativamente menores do que os arquivos de áudio não compactados, como CDs, permitindo fácil armazenamento e compartilhamento de música digital.

O MP3 teve um impacto profundo na indústria da música e na forma como consumimos música gerando um declínio das vendas de CDs: Os consumidores passaram a baixar e compartilhar músicas digitalmente. O MP3 impulsionou o crescimento da música digital, com o surgimento de serviços de streaming online como Spotify, Apple Music e Deezer. O MP3 abriu caminho para novos modelos de negócios na indústria da música, como a venda de músicas online e assinaturas de serviços de streaming.

Apesar de seu sucesso, o MP3 também enfrentou alguns desafios: Em taxas de bits mais baixas (como 64 kbps), a qualidade de som do MP3 pode ser perceptívelmente inferior à do áudio não compactado. Propriedades Intelectuais: Questões de direitos autorais e pirataria surgiram com a proliferação de arquivos MP3 compartilhados online. Formatos Alternativos: Novos formatos de áudio digital, como FLAC e AAC, oferecem maior qualidade de som ou taxas de compressão mais eficientes, mas ainda não alcançaram o mesmo nível de popularidade do MP3.

Outros padrões: MID (notação musical sintética), WAV (nenhuma perda de qualidade, logo arquivos grandes), além do formato OGG, da plataforma APPLE.

Vídeo No vídeo, estamos diante de necessidade maior de compressão. Se uma imagem estática já é grande, imagine precisar de 30 imagens estáticas por segundo. Agora, a localidade de referência ganha um novo viés, a l.r. no tempo. Ou seja, se imaginarmos um vídeo como um array 3-d (dimensões tempo, altura e largura), podemos comprimir em 2 dessas dimensões.

Aqui a história começa com os formatos analógicos (VHS, Betamax), segue pelo CD-ROM e VCD, e chega ao DVD. Depois vem os formatos AVC (padrão H.264), depois HEVC (H.265), VP9 (formato livre e aberto da Google) e AV1 (formato livre e aberto, melhor que o H.265). Diferentemente da imagem (e semelhantemente ao som), aqui existe muita preocupação com a latência (atraso no envio dos pacotes) que impacta diretamente a sincronia. O padrão H.264 utiliza timestamps e buffers para sincronizar áudio e vídeo. Em caso de perda de pacotes, mecanismos de recuperação de erros como o Concealment Motion Vector (CMV) podem ser utilizados para minimizar o impacto na sincronia. O CMV divide o vídeo em macroblocos. Quando há perda de um pacote contendo os vetores de movimento de um determinado macrobloco, o CMV busca os macroblocos vizinhos ($\uparrow, \downarrow, \leftarrow \rightarrow$) e ve se algum deles tem movimento similar ao esperado do macrobloco perdido. Daí ele "empresta" vetores e reconstitui o movimento ausente.

PDF O Portable Document Format (PDF), criado pela Adobe em 1993, se tornou um dos formatos de arquivo mais utilizados no mundo, presente em diversos setores e aplicações. Mais detalhes: vivx680a. Um documento PDF pode ser aberto e visualizado em qualquer dispositivo com um leitor de PDF instalado, independentemente do sistema operacional ou hardware utilizado. Isso garante que a formatação original do documento é sempre a mesma, independentemente da plataforma em que ele é aberto. Isso é crucial para documentos que exigem alta fidelidade visual, como apresentações, relatórios e publicações. O PDF oferece recursos de segurança robustos, como criptografia e assinaturas digitais, que protegem o conteúdo contra acessos não autorizados, modificações e falsificações. Isso torna o PDF ideal para documentos confidenciais e contratos legais. Até porque ele pode ser acompanhado de uma assinatura digital (*hash*). Suporta recursos de acessibilidade, como tags de estrutura e legendas para imagens, que facilitam o acesso ao conteúdo do documento por pessoas com deficiência visual ou outras necessidades especiais. Ao contrário de outros formatos de arquivo que podem ser corrompidos ou se tornar obsoletos com o tempo, o PDF é um formato durável e estável. Pense na economia de árvores: imagine se todos os PDFs existentes fossem para o papel. Ele é ideal para compartilhar documentos com outras pessoas, pois garante que a formatação original seja preservada e o conteúdo seja visualizado corretamente em diferentes dispositivos. O PDF é uma ótima opção para armazenar documentos importantes de forma segura e durável, pois protege o conteúdo contra modificações e garante que ele possa ser acessado no futuro. Revistas, livros, relatórios e outros materiais de publicação podem ser distribuídos no formato PDF, garantindo uma experiência de leitura consistente em diferentes dispositivos. **Outros:** Deixo de tratar aqui, por absoluta falta de espaço, outros BLOBs como exames médicos, arquivos de telescópios (vivx760), ...

Olhando dentro de um BLOB

Agora e hora de examinar com um "microscópio" as partes internas de um desses arquivos. O escolhido pela importância é o arquivo PDF. Pode parecer pouco, mas esta ideia simples, já salvou centenas de milhares (milhões ?) de árvores. Já são muitos os sistemas de informação que não geram mais saídas em papel, limitando-se a criar arquivos PDF.

Um arquivo PDF descreve como uma página ficará depois de impressa e ele é melhor do que imagem dessa página pelas seguintes razões

- ⋆ A imagem como tal sempre é muito maior do que sua descrição. A razão para isso é a evidente redundância de qualquer imagem.

⋆ A imagem como tal é fixa e só pode ser rotacionada, redimensionada ou de alguma maneira modificada à custa de processamento e perda de qualidade.

⋆ Sobre uma imagem não há como manipular o conteúdo, por exemplo conduzindo uma busca textual (^F alguma coisa ou então ^C e ^V).

⋆ Alterar uma única palavra sobre uma imagem exige habilidades de pintura. Sobre a descrição de uma página é tarefa simples: basta trocar uma palavra pela outra.

⋆ Imagine uma imagem enviada para um monitor (72 DPP) ou a mesma imagem enviada para uma impressora (300 DPP). Não podem ser a mesma imagem.

A versão 1.0 do padrão PDF nasceu em 1993 quando a Adobe lançou 2 programas: o Distiller (para gerar PDF) e o Reader (para ler): ambos pagos. A coisa começou a mudar quando o departamento de rendas americano comprou uma licença que permitia aos contribuintes baixar o Reader. Rapidamente a Adobe deu-se conta de que seria melhor para ela distribuir gratuitamente o Reader. Nos dez anos seguintes o PDF acabou se tornando o padrão de fato de descrição de imagens.

Apresenta as seguintes vantagens:

⋆ acesso aleatório: qualquer página pode ser montada independente das demais.

⋆ linearização: todos os elementos necessários para montar uma página estão juntos.

⋆ atualização incremental: mudanças ficam registradas no fim do arquivo. Vantagem: opção de desfazer ilimitada e rápida atualização. Desvantagem: o arquivo não pára de crescer.

⋆ fontes incorporadas: o destino sempre será montado corretamente. A fonte é desbastada de tudo que não é necessário.

⋆ padronização ISO: em 2008 a ISO abraçou o padrão Adobe PDF 1.7, o que o transformou em padrão aberto.

⋆ compatibilidade para trás e para a frente: Qualquer leitor lê as versões antigas e deve ler as futuras, já que todos os leitores ignoram o que não conhecem.

⋆ integrado a todos os principais browsers do protocolo `HTTP`.

Um arquivo PDF

Cabeçalho Formado por 2 linhas: a primeira identifica o arquivo como um PDF e informa a versão PDF em que ele foi gerado. A segunda linha inclui caracteres que não podem ser impressos (para avisar eventuais leitores deste fato). Exemplo:

```
%PDF-1.0
%âã
```

Corpo Um conjunto de nodos de um grafo, contendo as páginas, conteúdo gráfico, textos, sempre na forma de objetos. Cada objeto começa com um número de objeto (iniciando em 1), um número de geração (sempre 0) e a palavra chave `obj`. O objeto termina pela palavra chave `endobj`. Exemplo:

```
1 0 obj     -- descrevendo o objeto 1
<<          -- começa um dicionário
 /Kids [2 0 R]    -- /Kids tem o valor 2 0 R
 /Count 1         -- só um
 /Type /Pages     -- do tipo : página
>>               -- fim do dicionário
endobj            -- fim do objeto
```

Tabela de Referência Cruzada Lista o deslocamento em bytes de cada objeto no corpo do arquivo. Exemplo

```
0 6          -- seis entradas na tabela
0000000000  65535  f   -- entrada especial
0000000015  00000  n   -- obj1 no desloc 15
0000000074  00000  n   -- obj2 no desloc 74
...
```

Rodapé Informações e metadados para aumentar a performance de acesso. A primeira entrada é `trailer`. Depois o dicionário de trailer que deve apresentar ao menos a entrada `\Size` dizendo quantos itens há na tabela de referência cruzada e `\Root` dizendo qual objeto é o catálogo do documento. Depois vem uma linha que só contém `startxref` seguida por uma única linha indicando o deslocamento em bytes utilizado no inicio da tabela de referência cruzada do arquivo. Termina tudo a linha `%%EOF`. Exemplo:

```
trailer   -- palavra chave
<<   -- começa um dicionário
/Root 5 0 R   -- o catálogo é o objeto 5
/Size 6       -- tem o tamanho 6
>>   -- fim do dicionário
startxref
459  -- deslocamento da tabela de ref cruzada
%%EOF   -- fim de arquivo
```

Componentes

Uma descrição PDF suporta 5 componentes básicos: i. Números inteiros e reais (mas não exponenciais); ii. strings que são escritos entre parênteses; iii.Nomes, que sempre começam por uma barra normal; iv.Os valores booleanos `true` e `false` e v. O componente nulo (`null`). Aceita também 3 componentes compostos: i. Arrays, que são uma coleção ordenada de outros componentes, escritos entre colchetes; ii.Dicionários que são uma lista não ordenada de duplas: nome e componente. Começam por `<<` e terminam por `>>`. Finalmente iii. Fluxos, que armazenam dados binários, sempre acompanhados de um dicionário que descreve seus atributos, como comprimento e parâmetros de compactação, por exemplo.

☞ Para você fazer

As coisas em um arquivo PDF são separadas por um divisor de linhas. Pode ser o caracter $X'10'$, ou o $X'13'$ ou uma combinação de ambos. Ao examinar um PDF em hexadecimal, acostume-se a separar as linhas marcando este caractere. Isto posto, os componentes principais do arquivo PDF são:

- Header: identificação PDF e uma coleção de letras acentuadas
- Corpo: objetos numerados a partir de 1 sequencialmente, e identificados pela palavra `obj` no início e `endobj` no final. Esses objetos podem ser textos, desenhos, fontes, imagens, metadados, links, e um monte de coisas mais.
- Tabela de referência: uma lista de onde (no arquivo) está cada objeto. Esta tabela permite o acesso direto a cada um e a todos os objetos.
- Rodapé: Diversas informações que permitem o acesso rápido ao arquivo.

Nos interessa, neste exercício a informação `startxref` que é uma das últimas do arquivo. Você deve abrir o arquivo

```
arq10.pdf
```

publicado no lugar usual, com um editor de hexadecimal e deve procurar o startxref no final do arquivo. Na linha seguinte, existe um endereço que o PDF coloca em decimal. Você precisa transformar este número em hexadecimal e localizar no arquivo (mais acima) este endereço.

Nele, deve aparecer a palavra `xref` e na linha seguinte um zero e um número. Na linha seguinte, uma entrada (que todo arquivo tem) com muitos zeros, o número 65535 e uma letra. Na linha seguinte, **um número** seguido de zeros e uma letra. Trata-se do endereço do primeiro objeto do arquivo.

Responda aqui:

endereço da xref em hexadecimal	endereço do primeiro objeto em decimal

CEP - UP - UTFPR - PUCPr - UFPR –
17/06/2024 - 17:11:23.4
Prof Dr P Kantek (pkantek@gmail.com)
Objetos binários grandes vivxq10a, V: 1.01 76582
LUIZ MIGUEL OLINEK 24CC2301 - 10 entregar ate 4/julho _____ / _____ / _____

Objetos binários grandes - BLOBs

O objetivo desta atividade e conhecer e manusear os chamados objetos binários grandes: imagens, vídeos, documentos, músicas, etc. Tais objetos quando estão na memória do computador são conhecidos como BLOBs e quando vão para uma memória não volátil (pendrives, discos, SSDs, nuvem, ...) passam a ser arquivos.

Resultados da digitalização crescente da nossa sociedade, precisam ser conhecidos para deles se extraírem todos os benefícios que eles apresentam.

Eis alguns raciocínios: a documentação em papel de um avião boeing 747 ocuparia 5 aviões 747 para ser transportada. Alguém ainda compra enciclopédias em papel ? Lembram de quando as fotografias eram analógicas ? Um filme Ektachrome (36 fotos) custava bem caro. Vou chutar uns 150 reais, o que daria quase 5 reais/foto, só o filme, sem contar a revelação e as ampliações. Quando o I.R. era em papel, a receita precisava montar uma operação de guerra, envolvendo todos os bancos, apenas para receber as declarações, e isso que era mais ou menos 1/3 do número de contribuintes atual. E as músicas: durante muitos anos, no dia do meu pagamento eu ia a uma loja de discos (?) e comprava 1 único LP, ($\pm$ 30 minutos de música) era o que o meu dinheiro alcançava. Aliás, falando de música, vale comentar a ascensão e queda do formato CD. Ele passou de maravilha tecnológica em meados dos 80 a algo obsoleto em meados dos 2010. Do céu ao inferno em 30 anos. E, se perder em uma cidade estranha, tendo apenas o guia 4 Rodas para ajudar ? Uma sensação indescritível.

Vamos estudar mais de perto alguns BLOBs, lembrando que qualquer um pode criar e chamar de seu, um BLOB. Pela sua definição (grande bloco de dados binários com uma lógica interna de organização e acesso) vê-se que isto é perfeitamente possível. Assim, vamos nos limitar a BLOBs padrão como formatos e conceitos conhecidos e publicados.

Imagem É difícil contar esta história. Ela começa no daguerrótipo de meados do século 19, através da fixação de imagens em placas de vidro. Segue pelas imagens de TV que eram capturadas e transmitidas eletronicamente. Seu registro foi feito em gravações magnéticas analógicas. Uma nova necessidade surgiu na obtenção de fotos na Lua e além, quando não havia como retornar filmes. A propósito, um pouco antes disto, a vigilância eletrônica na Guerra Fria era muito custosa: satélites fotografavam o inimigo e precisavam ejetar os filmes (que acabavam rápido) ao sobrevoar território amigo. A primeira foto da lua, transmitida eletronicamente foi feita pela espaçonave Ranger em 1964, 17 minutos antes dela se espatifar na superfície lunar. Quando a Internet começou (por volta de 1990) a necessidade de manusear imagens sofreu uma mudança: antes o gerador e o visualizador eram de mesma origem, pelo que não havia necessidade de padronização. Com a Internet essa história mudou. Uma das primeiras respostas a esta demanda foi o formato BMP (Bitmap Image File). Embora originalmente proposto pela Microsoft e IBM, teve seu padrão publicado e virou de fato, um formato livre. Sua origem é a bio-física dos olhos humanos (3 canais separados para Red, Green e Blue), além de um truque bem legal chamado mapeamento pelo qual se negocia tamanho do arquivo VERSUS menor quantidade de cores. A principal inovação do BMP foi criar, publicar e obedecer a um padrão. Parece pouco, mas não é. A chave da popularização e do sucesso sempre é um PADRãO. Anote isso na sua cabeça e nunca mais esqueça. O padrão pode ser ruim ou bom (neste caso não é muito), mas é um padrão. A imagem é uma matriz numérica (outra inovação entusiasmadora), logo sujeita a sofrer a ação maravilhosa de qualquer matemática.

Padrão BMP O arquivo começa com um bloco de controle de 54 bytes, que contém o identificador 'BM', o tamanho do objeto (arquivo), onde a imagem começa, L=largura e A=altura da imagem, profundidade). Depois, dependendo da profundidade, pode haver uma tabela de cores. Finalmente a imagem na forma de $L \times A$ se for uma imagem mapeada ou 3 tabelas de $L \times A$, uma para cada cor se for uma imagem true-color. Veja na imagem a seguir estes conceitos

```
Exemplo:
      0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
      -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
0000  42 4D 62 05 00 00 00 00 00 00 36 04 00 00 28 00 BM.....6...(
0010  00 00 12 00 00 00 0F 00 00 00 01 00 08 00 00 00 ................
0020  00 00 60 01 00 00 00 00 00 00 00 00 00 00 00 01 ..`.............
0030  00 00 00 00 00 00 00 00 FF FF 00 80 80 80 00 FF ................
0040  FF 00 FF 00 FF 00 00 00 00 FF 00 00 00 00 00 FF ................
0050  00 00 FF FF 00 00 FF 00 00 00 09 09 09 00 0A 0A ................
0060  0A 00 0B 0B 0B 00 0C 0C 0C 00 0D 0D 0D 00 0E 0E ................
0070  0E 00 0F 0F 0F 00 10 10 10 00 11 11 11 00 12 12 ................
...
0410  F6 00 F7 F7 F7 00 F8 F8 F8 00 F9 F9 F9 00 FA FA ................
0420  FA 00 FB FB FB 00 FC FC FC 00 FD FD FD 00 FE FE ................
0430  FE 00 FF FF FF 00 01 01 03 01 01 01 08 01 01 01 ................
0440  01 01 02 01 01 06 01 01 00 00 04 04 03 04 04 04 ................
0450  04 04 04 04 04 04 04 04 04 04 04 04 00 00 01 01 ................
```

Os problemas do BMP incluem: nenhuma compressão, o que graças ao fenômeno da *localidade de referência espacial* é uma tragédia. Também não prevê nenhum tipo de animação, o que faz falta na Internet. Não sabia (hoje sabe) lidar com transparência.

Um concorrente importante do BMP (lá nos primórdios) foi o formato GIF que admite compressão (usando o belo algoritmo LZW) e também admite uma forma rudimentar de animação. O problema do GIF é que ele foi e é um formato proprietário. Isto deu origem ao formato PNG que é praticamente o mesmo, mas livre.

A compressão poderia ser muito melhor, se se admitisse algum nivel de degradação na reconstituição da imagem após sua transformação em BLOB. Isto foi conseguido no padrão JPEG, que ao admitir degradações (sob controle), gera compressões muito maiores do que GIF ou similares. Outra vantagem é a possibilidade de criação em etapas das imagens (coisa que o BMP nunca ofereceu). O padrão foi formalizado pela ISO em 1991, mas na época disparou uma briga por patentes e direitos. Nos EUA, as últimas patentes expiraram em 2004. A compressão aqui é pelo uso da Transformada Discreta do Coseno. A sua operação está além do nosso interesse (querendo, olhe VIVX690c ou 690p), mas em resumo: a imagem é quebrada em blocos de 8×8; esta matriz é multiplicada pela matriz DCT, o que passa do domínio do pixel para o domínio da frequência (Transformada de Fourier); os coeficientes são quantizados; os coeficientes são comprimidos. Depois, se faz o inverso e *voilà*.

Som O MP3, sigla para MPEG-1 Audio Layer III, é um formato de arquivo de áudio digital que revolucionou a forma como consumimos e compartilhamos música.

Desenvolvido no início da década de 1990 pelo Moving Picture Experts Group (MPEG), o MP3 se tornou o formato dominante para música digital, utilizado em players de música portáteis, computadores, smartphones e serviços de streaming online. O MP3 utiliza uma técnica de compressão com perda chamada percepção auditiva psicoacústica para reduzir o tamanho dos arquivos de áudio. Essa técnica remove informações da faixa de áudio que o ouvido humano provavelmente não consegue perceber, resultando em arquivos menores sem comprometer significativamente a qualidade da música para a maioria dos ouvintes. Apesar da compressão, o MP3 oferece uma qualidade de som bastante boa, especialmente em taxas de bits mais altas (como 128 kbps ou superior). Tamanho de Arquivo Reduzido: Os arquivos MP3 são significativamente menores do que os arquivos de áudio não compactados, como CDs, permitindo fácil armazenamento e compartilhamento de música digital.

O MP3 teve um impacto profundo na indústria da música e na forma como consumimos música gerando um declínio das vendas de CDs: Os consumidores passaram a baixar e compartilhar músicas digitalmente. O MP3 impulsionou o crescimento da música digital, com o surgimento de serviços de streaming online como Spotify, Apple Music e Deezer. O MP3 abriu caminho para novos modelos de negócios na indústria da música, como a venda de músicas online e assinaturas de serviços de streaming.

Apesar de seu sucesso, o MP3 também enfrentou alguns desafios: Em taxas de bits mais baixas (como 64 kbps), a qualidade de som do MP3 pode ser perceptívelmente inferior à do áudio não compactado. Propriedades Intelectuais: Questões de direitos autorais e pirataria surgiram com a proliferação de arquivos MP3 compartilhados online. Formatos Alternativos: Novos formatos de áudio digital, como FLAC e AAC, oferecem maior qualidade de som ou taxas de compressão mais eficientes, mas ainda não alcançaram o mesmo nível de popularidade do MP3.

Outros padrões: MID (notação musical sintética), WAV (nenhuma perda de qualidade, logo arquivos grandes), além do formato OGG, da plataforma APPLE.

Vídeo No vídeo, estamos diante de necessidade maior de compressão. Se uma imagem estática já é grande, imagine precisar de 30 imagens estáticas por segundo. Agora, a localidade de referência ganha um novo viés, a l.r. no tempo. Ou seja, se imaginarmos um vídeo como um array 3-d (dimensões tempo, altura e largura), podemos comprimir em 2 dessas dimensões.

Aqui a história começa com os formatos analógicos (VHS, Betamax), segue pelo CD-ROM e VCD, e chega ao DVD. Depois vem os formatos AVC (padrão H.264), depois HEVC (H.265), VP9 (formato livre e aberto da Google) e AV1 (formato livre e aberto, melhor que o H.265). Diferentemente da imagem (e semelhantemente ao som), aqui existe muita preocupação com a latência (atraso no envio dos pacotes) que impacta diretamente a sincronia. O padrão H.264 utiliza timestamps e buffers para sincronizar áudio e vídeo. Em caso de perda de pacotes, mecanismos de recuperação de erros como o Concealment Motion Vector (CMV) podem ser utilizados para minimizar o impacto na sincronia. O CMV divide o vídeo em macroblocos. Quando há perda de um pacote contendo os vetores de movimento de um determinado macrobloco, o CMV busca os macroblocos vizinhos ($\uparrow, \downarrow, \leftarrow \rightarrow$) e ve se algum deles tem movimento similar ao esperado do macrobloco perdido. Daí ele "empresta" vetores e reconstitui o movimento ausente.

PDF O Portable Document Format (PDF), criado pela Adobe em 1993, se tornou um dos formatos de arquivo mais utilizados no mundo, presente em diversos setores e aplicações. Mais detalhes: vivx680a. Um documento PDF pode ser aberto e visualizado em qualquer dispositivo com um leitor de PDF instalado, independentemente do sistema operacional ou hardware utilizado. Isso garante que a formatação original do documento é sempre a mesma, independentemente da plataforma em que ele é aberto. Isso é crucial para documentos que exigem alta fidelidade visual, como apresentações, relatórios e publicações. O PDF oferece recursos de segurança robustos, como criptografia e assinaturas digitais, que protegem o conteúdo contra acessos não autorizados, modificações e falsificações. Isso torna o PDF ideal para documentos confidenciais e contratos legais. Até porque ele pode ser acompanhado de uma assinatura digital (*hash*). Suporta recursos de acessibilidade, como tags de estrutura e legendas para imagens, que facilitam o acesso ao conteúdo do documento por pessoas com deficiência visual ou outras necessidades especiais. Ao contrário de outros formatos de arquivo que podem ser corrompidos ou se tornar obsoletos com o tempo, o PDF é um formato durável e estável. Pense na economia de árvores: imagine se todos os PDFs existentes fossem para o papel. Ele é ideal para compartilhar documentos com outras pessoas, pois garante que a formatação original seja preservada e o conteúdo seja visualizado corretamente em diferentes dispositivos. O PDF é uma ótima opção para armazenar documentos importantes de forma segura e durável, pois protege o conteúdo contra modificações e garante que ele possa ser acessado no futuro. Revistas, livros, relatórios e outros materiais de publicação podem ser distribuídos no formato PDF, garantindo uma experiência de leitura consistente em diferentes dispositivos. **Outros:** Deixo de tratar aqui, por absoluta falta de espaço, outros BLOBs como exames médicos, arquivos de telescópios (vivx760), ...

Olhando dentro de um BLOB

Agora e hora de examinar com um "microscópio" as partes internas de um desses arquivos. O escolhido pela importância é o arquivo PDF. Pode parecer pouco, mas esta ideia simples, já salvou centenas de milhares (milhões ?) de árvores. Já são muitos os sistemas de informação que não geram mais saídas em papel, limitando-se a criar arquivos PDF.

Um arquivo PDF descreve como uma página ficará depois de impressa e ele é melhor do que imagem dessa página pelas seguintes razões

- ⋆ A imagem como tal sempre é muito maior do que sua descrição. A razão para isso é a evidente redundância de qualquer imagem.

⋆ A imagem como tal é fixa e só pode ser rotacionada, redimensionada ou de alguma maneira modificada à custa de processamento e perda de qualidade.

⋆ Sobre uma imagem não há como manipular o conteúdo, por exemplo conduzindo uma busca textual (^F alguma coisa ou então ^C e ^V).

⋆ Alterar uma única palavra sobre uma imagem exige habilidades de pintura. Sobre a descrição de uma página é tarefa simples: basta trocar uma palavra pela outra.

⋆ Imagine uma imagem enviada para um monitor (72 DPP) ou a mesma imagem enviada para uma impressora (300 DPP). Não podem ser a mesma imagem.

A versão 1.0 do padrão PDF nasceu em 1993 quando a Adobe lançou 2 programas: o Distiller (para gerar PDF) e o Reader (para ler): ambos pagos. A coisa começou a mudar quando o departamento de rendas americano comprou uma licença que permitia aos contribuintes baixar o Reader. Rapidamente a Adobe deu-se conta de que seria melhor para ela distribuir gratuitamente o Reader. Nos dez anos seguintes o PDF acabou se tornando o padrão de fato de descrição de imagens.

Apresenta as seguintes vantagens:

⋆ acesso aleatório: qualquer página pode ser montada independente das demais.

⋆ linearização: todos os elementos necessários para montar uma página estão juntos.

⋆ atualização incremental: mudanças ficam registradas no fim do arquivo. Vantagem: opção de desfazer ilimitada e rápida atualização. Desvantagem: o arquivo não pára de crescer.

⋆ fontes incorporadas: o destino sempre será montado corretamente. A fonte é desbastada de tudo que não é necessário.

⋆ padronização ISO: em 2008 a ISO abraçou o padrão Adobe PDF 1.7, o que o transformou em padrão aberto.

⋆ compatibilidade para trás e para a frente: Qualquer leitor lê as versões antigas e deve ler as futuras, já que todos os leitores ignoram o que não conhecem.

⋆ integrado a todos os principais browsers do protocolo `HTTP`.

Um arquivo PDF

Cabeçalho Formado por 2 linhas: a primeira identifica o arquivo como um PDF e informa a versão PDF em que ele foi gerado. A segunda linha inclui caracteres que não podem ser impressos (para avisar eventuais leitores deste fato). Exemplo:

```
%PDF-1.0
%âã
```

Corpo Um conjunto de nodos de um grafo, contendo as páginas, conteúdo gráfico, textos, sempre na forma de objetos. Cada objeto começa com um número de objeto (iniciando em 1), um número de geração (sempre 0) e a palavra chave `obj`. O objeto termina pela palavra chave `endobj`. Exemplo:

```
1 0 obj     -- descrevendo o objeto 1
<<          -- começa um dicionário
 /Kids [2 0 R]     -- /Kids tem o valor 2 0 R
 /Count 1          -- só um
 /Type /Pages      -- do tipo : página
>>                 -- fim do dicionário
endobj             -- fim do objeto
```

Tabela de Referência Cruzada Lista o deslocamento em bytes de cada objeto no corpo do arquivo. Exemplo

```
0 6           -- seis entradas na tabela
0000000000  65535  f   -- entrada especial
0000000015  00000  n   -- obj1 no desloc 15
0000000074  00000  n   -- obj2 no desloc 74
...
```

Rodapé Informações e metadados para aumentar a performance de acesso. A primeira entrada é `trailer`. Depois o dicionário de trailer que deve apresentar ao menos a entrada `\Size` dizendo quantos itens há na tabela de referência cruzada e `\Root` dizendo qual objeto é o catálogo do documento. Depois vem uma linha que só contém `startxref` seguida por uma única linha indicando o deslocamento em bytes utilizado no inicio da tabela de referência cruzada do arquivo. Termina tudo a linha `%%EOF`. Exemplo:

```
trailer   -- palavra chave
<<   -- começa um dicionário
/Root 5 0 R   -- o catálogo é o objeto 5
/Size 6       -- tem o tamanho 6
>>   -- fim do dicionário
startxref
459  -- deslocamento da tabela de ref cruzada
%%EOF   -- fim de arquivo
```

Componentes

Uma descrição PDF suporta 5 componentes básicos: i. Números inteiros e reais (mas não exponenciais); ii. strings que são escritos entre parênteses; iii.Nomes, que sempre começam por uma barra normal; iv.Os valores booleanos `true` e `false` e v. O componente nulo (`null`). Aceita também 3 componentes compostos: i. Arrays, que são uma coleção ordenada de outros componentes, escritos entre colchetes; ii.Dicionários que são uma lista não ordenada de duplas: nome e componente. Começam por `<<` e terminam por `>>`. Finalmente iii. Fluxos, que armazenam dados binários, sempre acompanhados de um dicionário que descreve seus atributos, como comprimento e parâmetros de compactação, por exemplo.

☞ Para você fazer

As coisas em um arquivo PDF são separadas por um divisor de linhas. Pode ser o caracter $X'10'$, ou o $X'13'$ ou uma combinação de ambos. Ao examinar um PDF em hexadecimal, acostume-se a separar as linhas marcando este caractere. Isto posto, os componentes principais do arquivo PDF são:

- Header: identificação PDF e uma coleção de letras acentuadas
- Corpo: objetos numerados a partir de 1 sequencialmente, e identificados pela palavra `obj` no início e `endobj` no final. Esses objetos podem ser textos, desenhos, fontes, imagens, metadados, links, e um monte de coisas mais.
- Tabela de referência: uma lista de onde (no arquivo) está cada objeto. Esta tabela permite o acesso direto a cada um e a todos os objetos.
- Rodapé: Diversas informações que permitem o acesso rápido ao arquivo.

Nos interessa, neste exercício a informação `startxref` que é uma das últimas do arquivo. Você deve abrir o arquivo

```
arq02.pdf
```

publicado no lugar usual, com um editor de hexadecimal e deve procurar o startxref no final do arquivo. Na linha seguinte, existe um endereço que o PDF coloca em decimal. Você precisa transformar este número em hexadecimal e localizar no arquivo (mais acima) este endereço.

Nele, deve aparecer a palavra `xref` e na linha seguinte um zero e um número. Na linha seguinte, uma entrada (que todo arquivo tem) com muitos zeros, o número 65535 e uma letra. Na linha seguinte, **um número** seguido de zeros e uma letra. Trata-se do endereço do primeiro objeto do arquivo.

Responda aqui:

endereço da xref em hexadecimal	endereço do primeiro objeto em decimal

CEP - UP - UTFPR - PUCPr - UFPR –
17/06/2024 - 17:11:23.4
Prof Dr P Kantek (pkantek@gmail.com)
Objetos binários grandes vivxq10a, V: 1.01 76599
NICOLE DA SILVA DE AZEVEDO 24CC2301 -
11 entregar ate 4/julho _____ / _____ / _____

Objetos binários grandes - BLOBs

O objetivo desta atividade e conhecer e manusear os chamados objetos binários grandes: imagens, vídeos, documentos, músicas, etc. Tais objetos quando estão na memória do computador são conhecidos como BLOBs e quando vão para uma memória não volátil (pendrives, discos, SSDs, nuvem, ...) passam a ser arquivos.

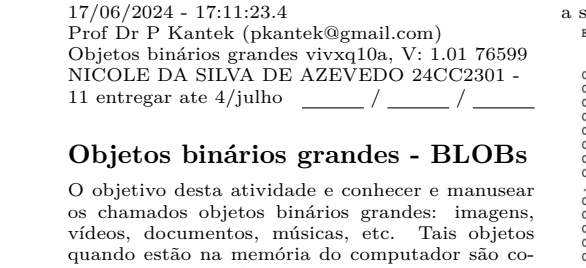
Resultados da digitalização crescente da nossa sociedade, precisam ser conhecidos para deles se extraírem todos os benefícios que eles apresentam.

Eis alguns raciocínios: a documentação em papel de um avião boeing 747 ocuparia 5 aviões 747 para ser transportada. Alguém ainda compra enciclopédias em papel ? Lembram de quando as fotografias eram analógicas ? Um filme Ektachrome (36 fotos) custava bem caro. Vou chutar uns 150 reais, o que daria quase 5 reais/foto, só o filme, sem contar a revelação e as ampliações. Quando o I.R. era em papel, a receita precisava montar uma operação de guerra, envolvendo todos os bancos, apenas para receber as declarações, e isso que era mais ou menos 1/3 do número de contribuintes atual. E as músicas: durante muitos anos, no dia do meu pagamento eu ia a uma loja de discos (?) e comprava 1 único LP, ($\pm$ 30 minutos de música) era o que o meu dinheiro alcançava. Aliás, falando de música, vale comentar a ascensão e queda do formato CD. Ele passou de maravilha tecnológica em meados dos 80 a algo obsoleto em meados dos 2010. Do céu ao inferno em 30 anos. E, se perder em uma cidade estranha, tendo apenas o guia 4 Rodas para ajudar ? Uma sensação indescritível.

Vamos estudar mais de perto alguns BLOBs, lembrando que qualquer um pode criar e chamar de seu, um BLOB. Pela sua definição (grande bloco de dados binários com uma lógica interna de organização e acesso) vê-se que isto é perfeitamente possível. Assim, vamos nos limitar a BLOBs padrão como formatos e conceitos conhecidos e publicados.

Imagem É difícil contar esta história. Ela começa no daguerrótipo de meados do século 19, através da fixação de imagens em placas de vidro. Segue pelas imagens de TV que eram capturadas e transmitidas eletronicamente. Seu registro foi feito em gravações magnéticas analógicas. Uma nova necessidade surgiu na obtenção de fotos na Lua e além, quando não havia como retornar filmes. A propósito, um pouco antes disto, a vigilância eletrônica na Guerra Fria era muito custosa: satélites fotografavam o inimigo e precisavam ejetar os filmes (que acabavam rápido) ao sobrevoar território amigo. A primeira foto da lua, transmitida eletronicamente foi feita pela espaçonave Ranger em 1964, 17 minutos antes dela se espatifar na superfície lunar. Quando a Internet começou (por volta de 1990) a necessidade de manusear imagens sofreu uma mudança: antes o gerador e o visualizador eram de mesma origem, pelo que não havia necessidade de padronização. Com a Internet essa história mudou. Uma das primeiras respostas a esta demanda foi o formato BMP (Bitmap Image File). Embora originalmente proposto pela Microsoft e IBM, teve seu padrão publicado e virou de fato, um formato livre. Sua origem é a bio-física dos olhos humanos (3 canais separados para Red, Green e Blue), além de um truque bem legal chamado mapeamento pelo qual se negocia tamanho do arquivo VERSUS menor quantidade de cores. A principal inovação do BMP foi criar, publicar e obedecer a um padrão. Parece pouco, mas não é. A chave da popularização e do sucesso sempre é um PADRãO. Anote isso na sua cabeça e nunca mais esqueça. O padrão pode ser ruim ou bom (neste caso não é muito), mas é um padrão. A imagem é uma matriz numérica (outra inovação entusiasmadora), logo sujeita a sofrer a ação maravilhosa de qualquer matemática.

Padrão BMP O arquivo começa com um bloco de controle de 54 bytes, que contém o identificador 'BM', o tamanho do objeto (arquivo), onde a imagem começa, L=largura e A=altura da imagem, profundidade). Depois, dependendo da profundidade, pode haver uma tabela de cores. Finalmente a imagem na forma de $L \times A$ se for uma imagem mapeada ou 3 tabelas de $L \times A$, uma para cada cor se for uma imagem true-color. Veja na imagem a seguir estes conceitos



Os problemas do BMP incluem: nenhuma compressão, o que graças ao fenômeno da *localidade de referência espacial* é uma tragédia. Também não prevê nenhum tipo de animação, o que faz falta na Internet. Não sabia (hoje sabe) lidar com transparência.

Um concorrente importante do BMP (lá nos primórdios) foi o formato GIF que admite compressão (usando o belo algoritmo LZW) e também admite uma forma rudimentar de animação. O problema do GIF é que ele foi e é um formato proprietário. Isto deu origem ao formato PNG que é praticamente o mesmo, mas livre.

A compressão poderia ser muito melhor, se se admitisse algum nivel de degradação na reconstituição da imagem após sua transformação em BLOB. Isto foi conseguido no padrão JPEG, que ao admitir degradações (sob controle), gera compressões muito maiores do que GIF ou similares. Outra vantagem é a possibilidade de criação em etapas das imagens (coisa que o BMP nunca ofereceu). O padrão foi formalizado pela ISO em 1991, mas na época disparou uma briga por patentes e direitos. Nos EUA, as últimas patentes expiraram em 2004. A compressão aqui é pelo uso da Transformada Discreta do Coseno. A sua operação está além do nosso interesse (querendo, olhe VIVX690c ou 690p), mas em resumo: a imagem é quebrada em blocos de 8×8; esta matriz é multiplicada pela matriz DCT, o que passa do domínio do pixel para o domínio da frequência (Transformada de Fourier); os coeficientes são quantizados; os coeficientes são comprimidos. Depois, se faz o inverso e *voilà*.

Som O MP3, sigla para MPEG-1 Audio Layer III, é um formato de arquivo de áudio digital que revolucionou a forma como consumimos e compartilhamos música.

Desenvolvido no início da década de 1990 pelo Moving Picture Experts Group (MPEG), o MP3 se tornou o formato dominante para música digital, utilizado em players de música portáteis, computadores, smartphones e serviços de streaming online. O MP3 utiliza uma técnica de compressão com perda chamada percepção auditiva psicoacústica para reduzir o tamanho dos arquivos de áudio. Essa técnica remove informações da faixa de áudio que o ouvido humano provavelmente não consegue perceber, resultando em arquivos menores sem comprometer significativamente a qualidade da música para a maioria dos ouvintes. Apesar da compressão, o MP3 oferece uma qualidade de som bastante boa, especialmente em taxas de bits mais altas (como 128 kbps ou superior). Tamanho de Arquivo Reduzido: Os arquivos MP3 são significativamente menores do que os arquivos de áudio não compactados, como CDs, permitindo fácil armazenamento e compartilhamento de música digital.

O MP3 teve um impacto profundo na indústria da música e na forma como consumimos música gerando um declínio das vendas de CDs: Os consumidores passaram a baixar e compartilhar músicas digitalmente. O MP3 impulsionou o crescimento da música digital, com o surgimento de serviços de streaming online como Spotify, Apple Music e Deezer. O MP3 abriu caminho para novos modelos de negócios na indústria da música, como a venda de músicas online e assinaturas de serviços de streaming.

Apesar de seu sucesso, o MP3 também enfrentou alguns desafios: Em taxas de bits mais baixas (como 64 kbps), a qualidade de som do MP3 pode ser perceptívelmente inferior à do áudio não compactado. Propriedades Intelectuais: Questões de direitos autorais e pirataria surgiram com a proliferação de arquivos MP3 compartilhados online. Formatos Alternativos: Novos formatos de áudio digital, como FLAC e AAC, oferecem maior qualidade de som ou taxas de compressão mais eficientes, mas ainda não alcançaram o mesmo nível de popularidade do MP3.

Outros padrões: MID (notação musical sintética), WAV (nenhuma perda de qualidade, logo arquivos grandes), além do formato OGG, da plataforma APPLE.

Vídeo No vídeo, estamos diante de necessidade maior de compressão. Se uma imagem estática já é grande, imagine precisar de 30 imagens estáticas por segundo. Agora, a localidade de referência ganha um novo viés, a l.r. no tempo. Ou seja, se imaginarmos um vídeo como um array 3-d (dimensões tempo, altura e largura), podemos comprimir em 2 dessas dimensões.

Aqui a história começa com os formatos analógicos (VHS, Betamax), segue pelo CD-ROM e VCD, e chega ao DVD. Depois vem os formatos AVC (padrão H.264), depois HEVC (H.265), VP9 (formato livre e aberto da Google) e AV1 (formato livre e aberto, melhor que o H.265). Diferentemente da imagem (e semelhantemente ao som), aqui existe muita preocupação com a latência (atraso no envio dos pacotes) que impacta diretamente a sincronia. O padrão H.264 utiliza timestamps e buffers para sincronizar áudio e vídeo. Em caso de perda de pacotes, mecanismos de recuperação de erros como o Concealment Motion Vector (CMV) podem ser utilizados para minimizar o impacto na sincronia. O CMV divide o vídeo em macroblocos. Quando há perda de um pacote contendo os vetores de movimento de um determinado macrobloco, o CMV busca os macroblocos vizinhos ($\uparrow, \downarrow, \leftarrow \rightarrow$) e ve se algum deles tem movimento similar ao esperado do macrobloco perdido. Daí ele "empresta" vetores e reconstitui o movimento ausente.

PDF O Portable Document Format (PDF), criado pela Adobe em 1993, se tornou um dos formatos de arquivo mais utilizados no mundo, presente em diversos setores e aplicações. Mais detalhes: vivx680a. Um documento PDF pode ser aberto e visualizado em qualquer dispositivo com um leitor de PDF instalado, independentemente do sistema operacional ou hardware utilizado. Isso garante que a formatação original do documento é sempre a mesma, independentemente da plataforma em que ele é aberto. Isso é crucial para documentos que exigem alta fidelidade visual, como apresentações, relatórios e publicações. O PDF oferece recursos de segurança robustos, como criptografia e assinaturas digitais, que protegem o conteúdo contra acessos não autorizados, modificações e falsificações. Isso torna o PDF ideal para documentos confidenciais e contratos legais. Até porque ele pode ser acompanhado de uma assinatura digital (*hash*). Suporta recursos de acessibilidade, como tags de estrutura e legendas para imagens, que facilitam o acesso ao conteúdo do documento por pessoas com deficiência visual ou outras necessidades especiais. Ao contrário de outros formatos de arquivo que podem ser corrompidos ou se tornar obsoletos com o tempo, o PDF é um formato durável e estável. Pense na economia de árvores: imagine se todos os PDFs existentes fossem para o papel. Ele é ideal para compartilhar documentos com outras pessoas, pois garante que a formatação original seja preservada e o conteúdo seja visualizado corretamente em diferentes dispositivos. O PDF é uma ótima opção para armazenar documentos importantes de forma segura e durável, pois protege o conteúdo contra modificações e garante que ele possa ser acessado no futuro. Revistas, livros, relatórios e outros materiais de publicação podem ser distribuídos no formato PDF, garantindo uma experiência de leitura consistente em diferentes dispositivos. **Outros:** Deixo de tratar aqui, por absoluta falta de espaço, outros BLOBs como exames médicos, arquivos de telescópios (vivx760), ...

Olhando dentro de um BLOB

Agora e hora de examinar com um "microscópio" as partes internas de um desses arquivos. O escolhido pela importância é o arquivo PDF. Pode parecer pouco, mas esta ideia simples, já salvou centenas de milhares (milhões ?) de árvores. Já são muitos os sistemas de informação que não geram mais saídas em papel, limitando-se a criar arquivos PDF.

Um arquivo PDF descreve como uma página ficará depois de impressa e ele é melhor do que imagem dessa página pelas seguintes razões

- ⋆ A imagem como tal sempre é muito maior do que sua descrição. A razão para isso é a evidente redundância de qualquer imagem.

⋆ A imagem como tal é fixa e só pode ser rotacionada, redimensionada ou de alguma maneira modificada à custa de processamento e perda de qualidade.

⋆ Sobre uma imagem não há como manipular o conteúdo, por exemplo conduzindo uma busca textual (^F alguma coisa ou então ^C e ^V).

⋆ Alterar uma única palavra sobre uma imagem exige habilidades de pintura. Sobre a descrição de uma página é tarefa simples: basta trocar uma palavra pela outra.

⋆ Imagine uma imagem enviada para um monitor (72 DPP) ou a mesma imagem enviada para uma impressora (300 DPP). Não podem ser a mesma imagem.

A versão 1.0 do padrão PDF nasceu em 1993 quando a Adobe lançou 2 programas: o Distiller (para gerar PDF) e o Reader (para ler): ambos pagos. A coisa começou a mudar quando o departamento de rendas americano comprou uma licença que permitia aos contribuintes baixar o Reader. Rapidamente a Adobe deu-se conta de que seria melhor para ela distribuir gratuitamente o Reader. Nos dez anos seguintes o PDF acabou se tornando o padrão de fato de descrição de imagens.

Apresenta as seguintes vantagens:

⋆ acesso aleatório: qualquer página pode ser montada independente das demais.

⋆ linearização: todos os elementos necessários para montar uma página estão juntos.

⋆ atualização incremental: mudanças ficam registradas no fim do arquivo. Vantagem: opção de desfazer ilimitada e rápida atualização. Desvantagem: o arquivo não pára de crescer.

⋆ fontes incorporadas: o destino sempre será montado corretamente. A fonte é desbastada de tudo que não é necessário.

⋆ padronização ISO: em 2008 a ISO abraçou o padrão Adobe PDF 1.7, o que o transformou em padrão aberto.

⋆ compatibilidade para trás e para a frente: Qualquer leitor lê as versões antigas e deve ler as futuras, já que todos os leitores ignoram o que não conhecem.

⋆ integrado a todos os principais browsers do protocolo `HTTP`.

Um arquivo PDF

Cabeçalho Formado por 2 linhas: a primeira identifica o arquivo como um PDF e informa a versão PDF em que ele foi gerado. A segunda linha inclui caracteres que não podem ser impressos (para avisar eventuais leitores deste fato). Exemplo:

```
%PDF-1.0
%âã
```

Corpo Um conjunto de nodos de um grafo, contendo as páginas, conteúdo gráfico, textos, sempre na forma de objetos. Cada objeto começa com um número de objeto (iniciando em 1), um número de geração (sempre 0) e a palavra chave `obj`. O objeto termina pela palavra chave `endobj`. Exemplo:

```
1 0 obj     -- descrevendo o objeto 1
<<          -- começa um dicionário
 /Kids [2 0 R]    -- /Kids tem o valor 2 0 R
 /Count 1         -- só um
 /Type /Pages     -- do tipo : página
>>               -- fim do dicionário
endobj            -- fim do objeto
```

Tabela de Referência Cruzada Lista o deslocamento em bytes de cada objeto no corpo do arquivo. Exemplo

```
0 6          -- seis entradas na tabela
0000000000  65535  f   -- entrada especial
0000000015  00000  n   -- obj1 no desloc 15
0000000074  00000  n   -- obj2 no desloc 74
...
```

Rodapé Informações e metadados para aumentar a performance de acesso. A primeira entrada é `trailer`. Depois o dicionário de trailer que deve apresentar ao menos a entrada `\Size` dizendo quantos itens há na tabela de referência cruzada e `\Root` dizendo qual objeto é o catálogo do documento. Depois vem uma linha que só contém `startxref` seguida por uma única linha indicando o deslocamento em bytes utilizado no inicio da tabela de referência cruzada do arquivo. Termina tudo a linha `%%EOF`. Exemplo:

```
trailer   -- palavra chave
<<   -- começa um dicionário
/Root 5 0 R   -- o catálogo é o objeto 5
/Size 6       -- tem o tamanho 6
>>   -- fim do dicionário
startxref
459  -- deslocamento da tabela de ref cruzada
%%EOF   -- fim de arquivo
```

Componentes

Uma descrição PDF suporta 5 componentes básicos: i. Números inteiros e reais (mas não exponenciais); ii. strings que são escritos entre parênteses; iii.Nomes, que sempre começam por uma barra normal; iv.Os valores booleanos `true` e `false` e v. O componente nulo (`null`). Aceita também 3 componentes compostos: i. Arrays, que são uma coleção ordenada de outros componentes, escritos entre colchetes; ii.Dicionários que são uma lista não ordenada de duplas: nome e componente. Começam por `<<` e terminam por `>>`. Finalmente iii. Fluxos, que armazenam dados binários, sempre acompanhados de um dicionário que descreve seus atributos, como comprimento e parâmetros de compactação, por exemplo.

☞ Para você fazer

As coisas em um arquivo PDF são separadas por um divisor de linhas. Pode ser o caracter $X'10'$, ou o $X'13'$ ou uma combinação de ambos. Ao examinar um PDF em hexadecimal, acostume-se a separar as linhas marcando este caractere. Isto posto, os componentes principais do arquivo PDF são:

- Header: identificação PDF e uma coleção de letras acentuadas
- Corpo: objetos numerados a partir de 1 sequencialmente, e identificados pela palavra `obj` no início e `endobj` no final. Esses objetos podem ser textos, desenhos, fontes, imagens, metadados, links, e um monte de coisas mais.
- Tabela de referência: uma lista de onde (no arquivo) está cada objeto. Esta tabela permite o acesso direto a cada um e a todos os objetos.
- Rodapé: Diversas informações que permitem o acesso rápido ao arquivo.

Nos interessa, neste exercício a informação `startxref` que é uma das últimas do arquivo. Você deve abrir o arquivo

```
arq06.pdf
```

publicado no lugar usual, com um editor de hexadecimal e deve procurar o startxref no final do arquivo. Na linha seguinte, existe um endereço que o PDF coloca em decimal. Você precisa transformar este número em hexadecimal e localizar no arquivo (mais acima) este endereço.

Nele, deve aparecer a palavra `xref` e na linha seguinte um zero e um número. Na linha seguinte, uma entrada (que todo arquivo tem) com muitos zeros, o número 65535 e uma letra. Na linha seguinte, **um número** seguido de zeros e uma letra. Trata-se do endereço do primeiro objeto do arquivo.

Responda aqui:

endereço da xref em hexadecimal	endereço do primeiro objeto em decimal

CEP - UP - UTFPR - PUCPr - UFPR –
17/06/2024 - 17:11:23.4
Prof Dr P Kantek (pkantek@gmail.com)
Objetos binários grandes vivxq10a, V: 1.01 76601
PAULO JOSE FREIRE CORREA 24CC2301 - 12
entregar ate 4/julho _____ / _____ / _____

Objetos binários grandes - BLOBs

O objetivo desta atividade e conhecer e manusear os chamados objetos binários grandes: imagens, vídeos, documentos, músicas, etc. Tais objetos quando estão na memória do computador são conhecidos como BLOBs e quando vão para uma memória não volátil (pendrives, discos, SSDs, nuvem, ...) passam a ser arquivos.

Resultados da digitalização crescente da nossa sociedade, precisam ser conhecidos para deles se extraírem todos os benefícios que eles apresentam.

Eis alguns raciocínios: a documentação em papel de um avião boeing 747 ocuparia 5 aviões 747 para ser transportada. Alguém ainda compra enciclopédias em papel ? Lembram de quando as fotografias eram analógicas ? Um filme Ektachrome (36 fotos) custava bem caro. Vou chutar uns 150 reais, o que daria quase 5 reais/foto, só o filme, sem contar a revelação e as ampliações. Quando o I.R. era em papel, a receita precisava montar uma operação de guerra, envolvendo todos os bancos, apenas para receber as declarações, e isso que era mais ou menos 1/3 do número de contribuintes atual. E as músicas: durante muitos anos, no dia do meu pagamento eu ia a uma loja de discos (?) e comprava 1 único LP, ($\pm$ 30 minutos de música) era o que o meu dinheiro alcançava. Aliás, falando de música, vale comentar a ascensão e queda do formato CD. Ele passou de maravilha tecnológica em meados dos 80 a algo obsoleto em meados dos 2010. Do céu ao inferno em 30 anos. E, se perder em uma cidade estranha, tendo apenas o guia 4 Rodas para ajudar ? Uma sensação indescritível.

Vamos estudar mais de perto alguns BLOBs, lembrando que qualquer um pode criar e chamar de seu, um BLOB. Pela sua definição (grande bloco de dados binários com uma lógica interna de organização e acesso) vê-se que isto é perfeitamente possível. Assim, vamos nos limitar a BLOBs padrão como formatos e conceitos conhecidos e publicados.

Imagem É difícil contar esta história. Ela começa no daguerrótipo de meados do século 19, através da fixação de imagens em placas de vidro. Segue pelas imagens de TV que eram capturadas e transmitidas eletronicamente. Seu registro foi feito em gravações magnéticas analógicas. Uma nova necessidade surgiu na obtenção de fotos na Lua e além, quando não havia como retornar filmes. A propósito, um pouco antes disto, a vigilância eletrônica na Guerra Fria era muito custosa: satélites fotografavam o inimigo e precisavam ejetar os filmes (que acabavam rápido) ao sobrevoar território amigo. A primeira foto da lua, transmitida eletronicamente foi feita pela espaçonave Ranger em 1964, 17 minutos antes dela se espatifar na superfície lunar. Quando a Internet começou (por volta de 1990) a necessidade de manusear imagens sofreu uma mudança: antes o gerador e o visualizador eram de mesma origem, pelo que não havia necessidade de padronização. Com a Internet essa história mudou. Uma das primeiras respostas a esta demanda foi o formato BMP (Bitmap Image File). Embora originalmente proposto pela Microsoft e IBM, teve seu padrão publicado e virou de fato, um formato livre. Sua origem é a bio-física dos olhos humanos (3 canais separados para Red, Green e Blue), além de um truque bem legal chamado mapeamento pelo qual se negocia tamanho do arquivo VERSUS menor quantidade de cores. A principal inovação do BMP foi criar, publicar e obedecer a um padrão. Parece pouco, mas não é. A chave da popularização e do sucesso sempre é um PADRãO. Anote isso na sua cabeça e nunca mais esqueça. O padrão pode ser ruim ou bom (neste caso não é muito), mas é um padrão. A imagem é uma matriz numérica (outra inovação entusiasmadora), logo sujeita a sofrer a ação maravilhosa de qualquer matemática.

Padrão BMP O arquivo começa com um bloco de controle de 54 bytes, que contém o identificador 'BM', o tamanho do objeto (arquivo), onde a imagem começa, L=largura e A=altura da imagem, profundidade). Depois, dependendo da profundidade, pode haver uma tabela de cores. Finalmente a imagem na forma de $L \times A$ se for uma imagem mapeada ou 3 tabelas de $L \times A$, uma para cada cor se for uma imagem true-color. Veja na imagem a seguir estes conceitos

Exemplo:

```
      0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
      -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
0000  42 4D 62 05 00 00 00 00 00 00 36 04 00 00 28 00 BM.....6...(
0010  00 00 12 00 00 00 0F 00 00 00 01 00 08 00 00 00 ................
0020  00 00 60 01 00 00 00 00 00 00 00 00 00 00 00 00 ..`.............
0030  00 00 00 00 00 00 00 00 FF FF 00 80 80 80 00 FF ................
0040  FF 00 FF 00 FF 00 00 00 00 FF 00 00 00 00 00 00 ................
0050  00 00 FF FF 00 00 FF 00 00 00 09 09 09 00 0A 0A ................
0060  0A 00 0B 0B 0B 00 0C 0C 0C 00 0D 0D 0D 00 0E 0E ................
0070  0E 00 0F 0F 0F 00 10 10 10 00 11 11 11 00 12 12 ................
...
0410  F6 00 F7 F7 F7 00 F8 F8 F8 00 F9 F9 F9 00 FA FA ................
0420  FA 00 FB FB FB 00 FC FC FC 00 FD FD FD 00 FE FE ................
0430  FE 00 FF FF FF 00 01 01 03 01 01 01 08 01 01 01 ................
0440  01 01 02 01 01 06 01 01 00 00 04 04 03 04 04 04 ................
0450  04 04 04 04 04 04 04 04 04 04 04 04 00 00 01 01 ................
```

Os problemas do BMP incluem: nenhuma compressão, o que graças ao fenômeno da *localidade de referência espacial* é uma tragédia. Também não prevê nenhum tipo de animação, o que faz falta na Internet. Não sabia (hoje sabe) lidar com transparência.

Um concorrente importante do BMP (lá nos primórdios) foi o formato GIF que admite compressão (usando o belo algoritmo LZW) e também admite uma forma rudimentar de animação. O problema do GIF é que ele foi e é um formato proprietário. Isto deu origem ao formato PNG que é praticamente o mesmo, mas livre.

A compressão poderia ser muito melhor, se se admitisse algum nivel de degradação na reconstituição da imagem após sua transformação em BLOB. Isto foi conseguido no padrão JPEG, que ao admitir degradações (sob controle), gera compressões muito maiores do que GIF ou similares. Outra vantagem é a possibilidade de criação em etapas das imagens (coisa que o BMP nunca ofereceu). O padrão foi formalizado pela ISO em 1991, mas na época disparou uma briga por patentes e direitos. Nos EUA, as últimas patentes expiraram em 2004. A compressão aqui é pelo uso da Transformada Discreta do Coseno. A sua operação está além do nosso interesse (querendo, olhe VIVX690c ou 690p), mas em resumo: a imagem é quebrada em blocos de 8×8; esta matriz é multiplicada pela matriz DCT, o que passa do domínio do pixel para o domínio da frequência (Transformada de Fourier); os coeficientes são quantizados; os coeficientes são comprimidos. Depois, se faz o inverso e *voilà*.

Som O MP3, sigla para MPEG-1 Audio Layer III, é um formato de arquivo de áudio digital que revolucionou a forma como consumimos e compartilhamos música.

Desenvolvido no início da década de 1990 pelo Moving Picture Experts Group (MPEG), o MP3 se tornou o formato dominante para música digital, utilizado em players de música portáteis, computadores, smartphones e serviços de streaming online. O MP3 utiliza uma técnica de compressão com perda chamada percepção auditiva psicoacústica para reduzir o tamanho dos arquivos de áudio. Essa técnica remove informações da faixa de áudio que o ouvido humano provavelmente não consegue perceber, resultando em arquivos menores sem comprometer significativamente a qualidade da música para a maioria dos ouvintes. Apesar da compressão, o MP3 oferece uma qualidade de som bastante boa, especialmente em taxas de bits mais altas (como 128 kbps ou superior). Tamanho de Arquivo Reduzido: Os arquivos MP3 são significativamente menores do que os arquivos de áudio não compactados, como CDs, permitindo fácil armazenamento e compartilhamento de música digital.

O MP3 teve um impacto profundo na indústria da música e na forma como consumimos música gerando um declínio das vendas de CDs: Os consumidores passaram a baixar e compartilhar músicas digitalmente. O MP3 impulsionou o crescimento da música digital, com o surgimento de serviços de streaming online como Spotify, Apple Music e Deezer. O MP3 abriu caminho para novos modelos de negócios na indústria da música, como a venda de músicas online e assinaturas de serviços de streaming.

Apesar de seu sucesso, o MP3 também enfrentou alguns desafios: Em taxas de bits mais baixas (como 64 kbps), a qualidade de som do MP3 pode ser perceptívelmente inferior à do áudio não compactado. Propriedades Intelectuais: Questões de direitos autorais e pirataria surgiram com a proliferação de arquivos MP3 compartilhados online. Formatos Alternativos: Novos formatos de áudio digital, como FLAC e AAC, oferecem maior qualidade de som ou taxas de compressão mais eficientes, mas ainda não alcançaram o mesmo nível de popularidade do MP3.

Outros padrões: MID (notação musical sintética), WAV (nenhuma perda de qualidade, logo arquivos grandes), além do formato OGG, da plataforma APPLE.

Vídeo No vídeo, estamos diante de necessidade maior de compressão. Se uma imagem estática já é grande, imagine precisar de 30 imagens estáticas por segundo. Agora, a localidade de referência ganha um novo viés, a l.r. no tempo. Ou seja, se imaginarmos um vídeo como um array 3-d (dimensões tempo, altura e largura), podemos comprimir em 2 dessas dimensões.

Aqui a história começa com os formatos analógicos (VHS, Betamax), segue pelo CD-ROM e VCD, e chega ao DVD. Depois vem os formatos AVC (padrão H.264), depois HEVC (H.265), VP9 (formato livre e aberto da Google) e AV1 (formato livre e aberto, melhor que o H.265). Diferentemente da imagem (e semelhantemente ao som), aqui existe muita preocupação com a latência (atraso no envio dos pacotes) que impacta diretamente a sincronia. O padrão H.264 utiliza timestamps e buffers para sincronizar áudio e vídeo. Em caso de perda de pacotes, mecanismos de recuperação de erros como o Concealment Motion Vector (CMV) podem ser utilizados para minimizar o impacto na sincronia. O CMV divide o vídeo em macroblocos. Quando há perda de um pacote contendo os vetores de movimento de um determinado macrobloco, o CMV busca os macroblocos vizinhos ($\uparrow, \downarrow, \leftarrow \rightarrow$) e ve se algum deles tem movimento similar ao esperado do macrobloco perdido. Daí ele "empresta" vetores e reconstitui o movimento ausente.

PDF O Portable Document Format (PDF), criado pela Adobe em 1993, se tornou um dos formatos de arquivo mais utilizados no mundo, presente em diversos setores e aplicações. Mais detalhes: vivx680a. Um documento PDF pode ser aberto e visualizado em qualquer dispositivo com um leitor de PDF instalado, independentemente do sistema operacional ou hardware utilizado. Isso garante que a formatação original do documento é sempre a mesma, independentemente da plataforma em que ele é aberto. Isso é crucial para documentos que exigem alta fidelidade visual, como apresentações, relatórios e publicações. O PDF oferece recursos de segurança robustos, como criptografia e assinaturas digitais, que protegem o conteúdo contra acessos não autorizados, modificações e falsificações. Isso torna o PDF ideal para documentos confidenciais e contratos legais. Até porque ele pode ser acompanhado de uma assinatura digital (*hash*). Suporta recursos de acessibilidade, como tags de estrutura e legendas para imagens, que facilitam o acesso ao conteúdo do documento por pessoas com deficiência visual ou outras necessidades especiais. Ao contrário de outros formatos de arquivo que podem ser corrompidos ou se tornar obsoletos com o tempo, o PDF é um formato durável e estável. Pense na economia de árvores: imagine se todos os PDFs existentes fossem para o papel. Ele é ideal para compartilhar documentos com outras pessoas, pois garante que a formatação original seja preservada e o conteúdo seja visualizado corretamente em diferentes dispositivos. O PDF é uma ótima opção para armazenar documentos importantes de forma segura e durável, pois protege o conteúdo contra modificações e garante que ele possa ser acessado no futuro. Revistas, livros, relatórios e outros materiais de publicação podem ser distribuídos no formato PDF, garantindo uma experiência de leitura consistente em diferentes dispositivos. **Outros:** Deixo de tratar aqui, por absoluta falta de espaço, outros BLOBs como exames médicos, arquivos de telescópios (vivx760), ...

Olhando dentro de um BLOB

Agora e hora de examinar com um "microscópio" as partes internas de um desses arquivos. O escolhido pela importância é o arquivo PDF. Pode parecer pouco, mas esta ideia simples, já salvou centenas de milhares (milhões ?) de árvores. Já são muitos os sistemas de informação que não geram mais saídas em papel, limitando-se a criar arquivos PDF.

Um arquivo PDF descreve como uma página ficará depois de impressa e ele é melhor do que imagem dessa página pelas seguintes razões

- ⋆ A imagem como tal sempre é muito maior do que sua descrição. A razão para isso é a evidente redundância de qualquer imagem.

⋆ A imagem como tal é fixa e só pode ser rotacionada, redimensionada ou de alguma maneira modificada à custa de processamento e perda de qualidade.

⋆ Sobre uma imagem não há como manipular o conteúdo, por exemplo conduzindo uma busca textual (^F alguma coisa ou então ^C e ^V).

⋆ Alterar uma única palavra sobre uma imagem exige habilidades de pintura. Sobre a descrição de uma página é tarefa simples: basta trocar uma palavra pela outra.

⋆ Imagine uma imagem enviada para um monitor (72 DPP) ou a mesma imagem enviada para uma impressora (300 DPP). Não podem ser a mesma imagem.

A versão 1.0 do padrão PDF nasceu em 1993 quando a Adobe lançou 2 programas: o Distiller (para gerar PDF) e o Reader (para ler): ambos pagos. A coisa começou a mudar quando o departamento de rendas americano comprou uma licença que permitia aos contribuintes baixar o Reader. Rapidamente a Adobe deu-se conta de que seria melhor para ela distribuir gratuitamente o Reader. Nos dez anos seguintes o PDF acabou se tornando o padrão de fato de descrição de imagens.

Apresenta as seguintes vantagens:

⋆ acesso aleatório: qualquer página pode ser montada independente das demais.

⋆ linearização: todos os elementos necessários para montar uma página estão juntos.

⋆ atualização incremental: mudanças ficam registradas no fim do arquivo. Vantagem: opção de desfazer ilimitada e rápida atualização. Desvantagem: o arquivo não pára de crescer.

⋆ fontes incorporadas: o destino sempre será montado corretamente. A fonte é desbastada de tudo que não é necessário.

⋆ padronização ISO: em 2008 a ISO abraçou o padrão Adobe PDF 1.7, o que o transformou em padrão aberto.

⋆ compatibilidade para trás e para a frente: Qualquer leitor lê as versões antigas e deve ler as futuras, já que todos os leitores ignoram o que não conhecem.

⋆ integrado a todos os principais browsers do protocolo `HTTP`.

Um arquivo PDF

Cabeçalho Formado por 2 linhas: a primeira identifica o arquivo como um PDF e informa a versão PDF em que ele foi gerado. A segunda linha inclui caracteres que não podem ser impressos (para avisar eventuais leitores deste fato). Exemplo:

```
%PDF-1.0
%âã
```

Corpo Um conjunto de nodos de um grafo, contendo as páginas, conteúdo gráfico, textos, sempre na forma de objetos. Cada objeto começa com um número de objeto (iniciando em 1), um número de geração (sempre 0) e a palavra chave `obj`. O objeto termina pela palavra chave `endobj`. Exemplo:

```
1 0 obj     -- descrevendo o objeto 1
<<          -- começa um dicionário
 /Kids [2 0 R]    -- /Kids tem o valor 2 0 R
 /Count 1         -- só um
 /Type /Pages     -- do tipo : página
>>               -- fim do dicionário
endobj            -- fim do objeto
```

Tabela de Referência Cruzada Lista o deslocamento em bytes de cada objeto no corpo do arquivo. Exemplo

```
0 6          -- seis entradas na tabela
0000000000  65535  f   -- entrada especial
0000000015  00000  n   -- obj1 no desloc 15
0000000074  00000  n   -- obj2 no desloc 74
...
```

Rodapé Informações e metadados para aumentar a performance de acesso. A primeira entrada é `trailer`. Depois o dicionário de trailer que deve apresentar ao menos a entrada `\Size` dizendo quantos itens há na tabela de referência cruzada e `\Root` dizendo qual objeto é o catálogo do documento. Depois vem uma linha que só contém `startxref` seguida por uma única linha indicando o deslocamento em bytes utilizado no inicio da tabela de referência cruzada do arquivo. Termina tudo a linha `%%EOF`. Exemplo:

```
trailer   -- palavra chave
<<   -- começa um dicionário
/Root 5 0 R   -- o catálogo é o objeto 5
/Size 6       -- tem o tamanho 6
>>   -- fim do dicionário
startxref
459  -- deslocamento da tabela de ref cruzada
%%EOF   -- fim de arquivo
```

Componentes

Uma descrição PDF suporta 5 componentes básicos: i. Números inteiros e reais (mas não exponenciais); ii. strings que são escritos entre parênteses; iii.Nomes, que sempre começam por uma barra normal; iv.Os valores booleanos `true` e `false` e v. O componente nulo (`null`). Aceita também 3 componentes compostos: i. Arrays, que são uma coleção ordenada de outros componentes, escritos entre colchetes; ii.Dicionários que são uma lista não ordenada de duplas: nome e componente. Começam por `<<` e terminam por `>>`. Finalmente iii. Fluxos, que armazenam dados binários, sempre acompanhados de um dicionário que descreve seus atributos, como comprimento e parâmetros de compactação, por exemplo.

☞ Para você fazer

As coisas em um arquivo PDF são separadas por um divisor de linhas. Pode ser o caracter $X'10'$, ou o $X'13'$ ou uma combinação de ambos. Ao examinar um PDF em hexadecimal, acostume-se a separar as linhas marcando este caractere. Isto posto, os componentes principais do arquivo PDF são:

- Header: identificação PDF e uma coleção de letras acentuadas
- Corpo: objetos numerados a partir de 1 sequencialmente, e identificados pela palavra `obj` no início e `endobj` no final. Esses objetos podem ser textos, desenhos, fontes, imagens, metadados, links, e um monte de coisas mais.
- Tabela de referência: uma lista de onde (no arquivo) está cada objeto. Esta tabela permite o acesso direto a cada um e a todos os objetos.
- Rodapé: Diversas informações que permitem o acesso rápido ao arquivo.

Nos interessa, neste exercício a informação `startxref` que é uma das últimas do arquivo. Você deve abrir o arquivo

```
arq09.pdf
```

publicado no lugar usual, com um editor de hexadecimal e deve procurar o startxref no final do arquivo. Na linha seguinte, existe um endereço que o PDF coloca em decimal. Você precisa transformar este número em hexadecimal e localizar no arquivo (mais acima) este endereço.

Nele, deve aparecer a palavra `xref` e na linha seguinte um zero e um número. Na linha seguinte, uma entrada (que todo arquivo tem) com muitos zeros, o número 65535 e uma letra. Na linha seguinte, **um número** seguido de zeros e uma letra. Trata-se do endereço do primeiro objeto do arquivo.

Responda aqui:

endereço da xref em hexadecimal	endereço do primeiro objeto em decimal

CEP - UP - UTFPR - PUCPr - UFPR –
17/06/2024 - 17:11:23.4
Prof Dr P Kantek (pkantek@gmail.com)
Objetos binários grandes vivxq10a, V: 1.01 76618
PEDRO HENRIQUE PIEKARSKI DE OL
24CC2301 - 13 entregar ate 4/julho _____ /
/ _____

Objetos binários grandes - BLOBs

O objetivo desta atividade e conhecer e manusear os chamados objetos binários grandes: imagens, vídeos, documentos, músicas, etc. Tais objetos quando estão na memória do computador são conhecidos como BLOBs e quando vão para uma memória não volátil (pendrives, discos, SSDs, nuvem, ...) passam a ser arquivos.

Resultados da digitalização crescente da nossa sociedade, precisam ser conhecidos para deles se extraírem todos os benefícios que eles apresentam.

Eis alguns raciocínios: a documentação em papel de um avião boeing 747 ocuparia 5 aviões 747 para ser transportada. Alguém ainda compra enciclopédias em papel ? Lembram de quando as fotografias eram analógicas ? Um filme Ektachrome (36 fotos) custava bem caro. Vou chutar uns 150 reais, o que daria quase 5 reais/foto, só o filme, sem contar a revelação e as ampliações. Quando o I.R. era em papel, a receita precisava montar uma operação de guerra, envolvendo todos os bancos, apenas para receber as declarações, e isso que era mais ou menos 1/3 do número de contribuintes atual. E as músicas: durante muitos anos, no dia do meu pagamento eu ia a uma loja de discos (?) e comprava 1 único LP, ($\pm$ 30 minutos de música) era o que o meu dinheiro alcançava. Aliás, falando de música, vale comentar a ascensão e queda do formato CD. Ele passou de maravilha tecnológica em meados dos 80 a algo obsoleto em meados dos 2010. Do céu ao inferno em 30 anos. E, se perder em uma cidade estranha, tendo apenas o guia 4 Rodas para ajudar ? Uma sensação indescritível.

Vamos estudar mais de perto alguns BLOBs, lembrando que qualquer um pode criar e chamar de seu, um BLOB. Pela sua definição (grande bloco de dados binários com uma lógica interna de organização e acesso) vê-se que isto é perfeitamente possível. Assim, vamos nos limitar a BLOBs padrão como formatos e conceitos conhecidos e publicados.

Imagem É difícil contar esta história. Ela começa no daguerrótipo de meados do século 19, através da fixação de imagens em placas de vidro. Segue pelas imagens de TV que eram capturadas e transmitidas eletronicamente. Seu registro foi feito em gravações magnéticas analógicas. Uma nova necessidade surgiu na obtenção de fotos na Lua e além, quando não havia como retornar filmes. A propósito, um pouco antes disto, a vigilância eletrônica na Guerra Fria era muito custosa: satélites fotografavam o inimigo e precisavam ejetar os filmes (que acabavam rápido) ao sobrevoar território amigo. A primeira foto da lua, transmitida eletronicamente foi feita pela espaçonave Ranger em 1964, 17 minutos antes dela se espatifar na superfície lunar. Quando a Internet começou (por volta de 1990) a necessidade de manusear imagens sofreu uma mudança: antes o gerador e o visualizador eram de mesma origem, pelo que não havia necessidade de padronização. Com a Internet essa história mudou. Uma das primeiras respostas a esta demanda foi o formato BMP (Bitmap Image File). Embora originalmente proposto pela Microsoft e IBM, teve seu padrão publicado e virou de fato, um formato livre. Sua origem é a bio-física dos olhos humanos (3 canais separados para Red, Green e Blue), além de um truque bem legal chamado mapeamento pelo qual se negocia tamanho do arquivo VERSUS menor quantidade de cores. A principal inovação do BMP foi criar, publicar e obedecer a um padrão. Parece pouco, mas não é. A chave da popularização e do sucesso sempre é um PADRãO. Anote isso na sua cabeça e nunca mais esqueça. O padrão pode ser ruim ou bom (neste caso não é muito), mas é um padrão. A imagem é uma matriz numérica (outra inovação entusiasmadora), logo sujeita a sofrer a ação maravilhosa de qualquer matemática.

Padrão BMP O arquivo começa com um bloco de controle de 54 bytes, que contém o identificador 'BM', o tamanho do objeto (arquivo), onde a imagem começa, L=largura e A=altura da imagem, profundidade). Depois, dependendo da profundidade, pode haver uma tabela de cores. Finalmente a imagem na forma de $L \times A$ se for uma imagem mapeada ou 3 tabelas de $L \times A$, uma para cada cor se for uma imagem true-color. Veja na imagem a seguir estes conceitos

```
Exemplo:
        0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
        -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
0000  42 4D 62 05 00 00 00 00 00 00 36 04 00 00 28 00 BM.....6...(
0010  00 00 12 00 00 00 0F 00 00 00 01 00 08 00 00 00 ................
0020  00 00 60 01 00 00 00 00 00 00 00 00 00 00 00 01 ................
0030  00 00 00 00 00 00 00 00 FF FF 00 80 80 80 00 FF FF ................
0040  FF 00 FF 00 FF 00 00 00 FF 00 00 00 00 00 00 FF ................
0050  00 00 FF FF 00 00 FF 00 00 00 09 09 09 00 0A 0A ................
0060  0A 00 0B 0B 0B 00 0C 0C 0C 00 0D 0D 0D 00 0E 0E ................
0070  0E 00 0F 0F 0F 00 10 10 10 00 11 11 11 00 12 12 ................
...
0410  F6 00 F7 F7 F7 00 F8 F8 F8 00 F9 F9 F9 00 FA FA ................
0420  FA 00 FB FB FB 00 FC FC FC 00 FD FD FD 00 FE FE ................
0430  FE 00 FF FF FF 00 01 01 03 01 01 01 08 01 01 01 ................
0440  01 01 02 01 01 06 01 01 00 00 04 04 03 04 04 04 ................
0450  04 04 04 04 04 04 04 04 04 04 04 04 00 00 01 01 ................
```

Os problemas do BMP incluem: nenhuma compressão, o que graças ao fenômeno da *localidade de referência espacial* é uma tragédia. Também não prevê nenhum tipo de animação, o que faz falta na Internet. Não sabia (hoje sabe) lidar com transparência.

Um concorrente importante do BMP (lá nos primórdios) foi o formato GIF que admite compressão (usando o belo algoritmo LZW) e também admite uma forma rudimentar de animação. O problema do GIF é que ele foi e é um formato proprietário. Isto deu origem ao formato PNG que é praticamente o mesmo, mas livre.

A compressão poderia ser muito melhor, se se admitisse algum nivel de degradação na reconstituição da imagem após sua transformação em BLOB. Isto foi conseguido no padrão JPEG, que ao admitir degradações (sob controle), gera compressões muito maiores do que GIF ou similares. Outra vantagem é a possibilidade de criação em etapas das imagens (coisa que o BMP nunca ofereceu). O padrão foi formalizado pela ISO em 1991, mas na época disparou uma briga por patentes e direitos. Nos EUA, as últimas patentes expiraram em 2004. A compressão aqui é pelo uso da Transformada Discreta do Coseno. A sua operação está além do nosso interesse (querendo, olhe VIVX690c ou 690p), mas em resumo: a imagem é quebrada em blocos de 8×8; esta matriz é multiplicada pela matriz DCT, o que passa do domínio do pixel para o domínio da frequência (Transformada de Fourier); os coeficientes são quantizados; os coeficientes são comprimidos. Depois, se faz o inverso e *voilà*.

Som O MP3, sigla para MPEG-1 Audio Layer III, é um formato de arquivo de áudio digital que revolucionou a forma como consumimos e compartilhamos música.

Desenvolvido no início da década de 1990 pelo Moving Picture Experts Group (MPEG), o MP3 se tornou o formato dominante para música digital, utilizado em players de música portáteis, computadores, smartphones e serviços de streaming online. O MP3 utiliza uma técnica de compressão com perda chamada percepção auditiva psicoacústica para reduzir o tamanho dos arquivos de áudio. Essa técnica remove informações da faixa de áudio que o ouvido humano provavelmente não consegue perceber, resultando em arquivos menores sem comprometer significativamente a qualidade da música para a maioria dos ouvintes. Apesar da compressão, o MP3 oferece uma qualidade de som bastante boa, especialmente em taxas de bits mais altas (como 128 kbps ou superior). Tamanho de Arquivo Reduzido: Os arquivos MP3 são significativamente menores do que os arquivos de áudio não compactados, como CDs, permitindo fácil armazenamento e compartilhamento de música digital.

O MP3 teve um impacto profundo na indústria da música e na forma como consumimos música gerando um declínio das vendas de CDs: Os consumidores passaram a baixar e compartilhar músicas digitalmente. O MP3 impulsionou o crescimento da música digital, com o surgimento de serviços de streaming online como Spotify, Apple Music e Deezer. O MP3 abriu caminho para novos modelos de negócios na indústria da música, como a venda de músicas online e assinaturas de serviços de streaming.

Apesar de seu sucesso, o MP3 também enfrentou alguns desafios: Em taxas de bits mais baixas (como 64 kbps), a qualidade de som do MP3 pode ser perceptívelmente inferior à do áudio não compactado. Propriedades Intelectuais: Questões de direitos autorais e pirataria surgiram com a proliferação de arquivos MP3 compartilhados online. Formatos Alternativos: Novos formatos de áudio digital, como FLAC e AAC, oferecem maior qualidade de som ou taxas de compressão mais eficientes, mas ainda não alcançaram o mesmo nível de popularidade do MP3.

Outros padrões: MID (notação musical sintética), WAV (nenhuma perda de qualidade, logo arquivos grandes), além do formato OGG, da plataforma APPLE.

Vídeo No vídeo, estamos diante de necessidade maior de compressão. Se uma imagem estática já é grande, imagine precisar de 30 imagens estáticas por segundo. Agora, a localidade de referência ganha um novo viés, a l.r. no tempo. Ou seja, se imaginarmos um vídeo como um array 3-d (dimensões tempo, altura e largura), podemos comprimir em 2 dessas dimensões.

Aqui a história começa com os formatos analógicos (VHS, Betamax), segue pelo CD-ROM e VCD, e chega ao DVD. Depois vem os formatos AVC (padrão H.264), depois HEVC (H.265), VP9 (formato livre e aberto da Google) e AV1 (formato livre e aberto, melhor que o H.265). Diferentemente da imagem (e semelhantemente ao som), aqui existe muita preocupação com a latência (atraso no envio dos pacotes) que impacta diretamente a sincronia. O padrão H.264 utiliza timestamps e buffers para sincronizar áudio e vídeo. Em caso de perda de pacotes, mecanismos de recuperação de erros como o Concealment Motion Vector (CMV) podem ser utilizados para minimizar o impacto na sincronia. O CMV divide o vídeo em macroblocos. Quando há perda de um pacote contendo os vetores de movimento de um determinado macrobloco, o CMV busca os macroblocos vizinhos ($\uparrow, \downarrow, \leftarrow \rightarrow$) e ve se algum deles tem movimento similar ao esperado do macrobloco perdido. Daí ele "empresta" vetores e reconstitui o movimento ausente.

PDF O Portable Document Format (PDF), criado pela Adobe em 1993, se tornou um dos formatos de arquivo mais utilizados no mundo, presente em diversos setores e aplicações. Mais detalhes: vivx680a. Um documento PDF pode ser aberto e visualizado em qualquer dispositivo com um leitor de PDF instalado, independentemente do sistema operacional ou hardware utilizado. Isso garante que a formatação original do documento é sempre a mesma, independentemente da plataforma em que ele é aberto. Isso é crucial para documentos que exigem alta fidelidade visual, como apresentações, relatórios e publicações. O PDF oferece recursos de segurança robustos, como criptografia e assinaturas digitais, que protegem o conteúdo contra acessos não autorizados, modificações e falsificações. Isso torna o PDF ideal para documentos confidenciais e contratos legais. Até porque ele pode ser acompanhado de uma assinatura digital (*hash*). Suporta recursos de acessibilidade, como tags de estrutura e legendas para imagens, que facilitam o acesso ao conteúdo do documento por pessoas com deficiência visual ou outras necessidades especiais. Ao contrário de outros formatos de arquivo que podem ser corrompidos ou se tornar obsoletos com o tempo, o PDF é um formato durável e estável. Pense na economia de árvores: imagine se todos os PDFs existentes fossem para o papel. Ele é ideal para compartilhar documentos com outras pessoas, pois garante que a formatação original seja preservada e o conteúdo seja visualizado corretamente em diferentes dispositivos. O PDF é uma ótima opção para armazenar documentos importantes de forma segura e durável, pois protege o conteúdo contra modificações e garante que ele possa ser acessado no futuro. Revistas, livros, relatórios e outros materiais de publicação podem ser distribuídos no formato PDF, garantindo uma experiência de leitura consistente em diferentes dispositivos. **Outros:** Deixo de tratar aqui, por absoluta falta de espaço, outros BLOBs como exames médicos, arquivos de telescópios (vivx760), ...

Olhando dentro de um BLOB

Agora e hora de examinar com um "microscópio" as partes internas de um desses arquivos. O escolhido pela importância é o arquivo PDF. Pode parecer pouco, mas esta ideia simples, já salvou centenas de milhares (milhões ?) de árvores. Já são muitos os sistemas de informação que não geram mais saídas em papel, limitando-se a criar arquivos PDF.

Um arquivo PDF descreve como uma página ficará depois de impressa e ele é melhor do que imagem dessa página pelas seguintes razões

⋆ A imagem como tal sempre é muito maior do que

sua descrição. A razão para isso é a evidente redundância de qualquer imagem.

⋆ A imagem como tal é fixa e só pode ser rotacionada, redimensionada ou de alguma maneira modificada à custa de processamento e perda de qualidade.

⋆ Sobre uma imagem não há como manipular o conteúdo, por exemplo conduzindo uma busca textual (^F alguma coisa ou então ^C e ^V).

⋆ Alterar uma única palavra sobre uma imagem exige habilidades de pintura. Sobre a descrição de uma página é tarefa simples: basta trocar uma palavra pela outra.

⋆ Imagine uma imagem enviada para um monitor (72 DPP) ou a mesma imagem enviada para uma impressora (300 DPP). Não podem ser a mesma imagem.

A versão 1.0 do padrão PDF nasceu em 1993 quando a Adobe lançou 2 programas: o Distiller (para gerar PDF) e o Reader (para ler): ambos pagos. A coisa começou a mudar quando o departamento de rendas americano comprou uma licença que permitia aos contribuintes baixar o Reader. Rapidamente a Adobe deu-se conta de que seria melhor para ela distribuir gratuitamente o Reader. Nos dez anos seguintes o PDF acabou se tornando o padrão de fato de descrição de imagens.

Apresenta as seguintes vantagens:

⋆ acesso aleatório: qualquer página pode ser montada independente das demais.

⋆ linearização: todos os elementos necessários para montar uma página estão juntos.

⋆ atualização incremental: mudanças ficam registradas no fim do arquivo. Vantagem: opção de desfazer ilimitada e rápida atualização. Desvantagem: o arquivo não pára de crescer.

⋆ fontes incorporadas: o destino sempre será montado corretamente. A fonte é desbastada de tudo que não é necessário.

⋆ padronização ISO: em 2008 a ISO abraçou o padrão Adobe PDF 1.7, o que o transformou em padrão aberto.

⋆ compatibilidade para trás e para a frente: Qualquer leitor lê as versões antigas e deve ler as futuras, já que todos os leitores ignoram o que não conhecem.

⋆ integrado a todos os principais browsers do protocolo `HTTP`.

Um arquivo PDF

Cabeçalho Formado por 2 linhas: a primeira identifica o arquivo como um PDF e informa a versão PDF em que ele foi gerado. A segunda linha inclui caracteres que não podem ser impressos (para avisar eventuais leitores deste fato). Exemplo:

```
%PDF-1.0
%âã
```

Corpo Um conjunto de nodos de um grafo, contendo as páginas, conteúdo gráfico, textos, sempre na forma de objetos. Cada objeto começa com um número de objeto (iniciando em 1), um número de geração (sempre 0) e a palavra chave `obj`. O objeto termina pela palavra chave `endobj`. Exemplo:

```
1 0 obj     -- descrevendo o objeto 1
<<          -- começa um dicionário
 /Kids [2 0 R]     -- /Kids tem o valor 2 0 R
 /Count 1          -- só um
 /Type /Pages      -- do tipo : página
>>                 -- fim do dicionário
endobj             -- fim do objeto
```

Tabela de Referência Cruzada Lista o deslocamento em bytes de cada objeto no corpo do arquivo. Exemplo

```
0 6           -- seis entradas na tabela
0000000000  65535  f   -- entrada especial
0000000015  00000  n   -- obj1 no desloc 15
0000000074  00000  n   -- obj2 no desloc 74
...
```

Rodapé Informações e metadados para aumentar a performance de acesso. A primeira entrada é `trailer`. Depois o dicionário de trailer que deve apresentar ao menos a entrada `\Size` dizendo quantos itens há na tabela de referência cruzada e `\Root` dizendo qual objeto é o catálogo do documento. Depois vem uma linha que só contém `startxref` seguida por uma única linha indicando o deslocamento em bytes utilizado no inicio da tabela de referência cruzada do arquivo. Termina tudo a linha `%%EOF`. Exemplo:

```
trailer   -- palavra chave
<<   -- começa um dicionário
/Root 5 0 R   -- o catálogo é o objeto 5
/Size 6       -- tem o tamanho 6
>>   -- fim do dicionário
startxref
459  -- deslocamento da tabela de ref cruzada
%%EOF   -- fim de arquivo
```

Componentes

Uma descrição PDF suporta 5 componentes básicos: i. Números inteiros e reais (mas não exponenciais); ii. strings que são escritos entre parênteses; iii.Nomes, que sempre começam por uma barra normal; iv.Os valores booleanos `true` e `false` e v. O componente nulo (`null`). Aceita também 3 componentes compostos: i. Arrays, que são uma coleção ordenada de outros componentes, escritos entre colchetes; ii.Dicionários que são uma lista não ordenada de duplas: nome e componente. Começam por `<<` e terminam por `>>`. Finalmente iii. Fluxos, que armazenam dados binários, sempre acompanhados de um dicionário que descreve seus atributos, como comprimento e parâmetros de compactação, por exemplo.

☞ Para você fazer

As coisas em um arquivo PDF são separadas por um divisor de linhas. Pode ser o caracter $X'10'$, ou o $X'13'$ ou uma combinação de ambos. Ao examinar um PDF em hexadecimal, acostume-se a separar as linhas marcando este caractere. Isto posto, os componentes principais do arquivo PDF são:

- Header: identificação PDF e uma coleção de letras acentuadas
- Corpo: objetos numerados a partir de 1 sequencialmente, e identificados pela palavra `obj` no início e `endobj` no final. Esses objetos podem ser textos, desenhos, fontes, imagens, metadados, links, e um monte de coisas mais.
- Tabela de referência: uma lista de onde (no arquivo) está cada objeto. Esta tabela permite o acesso direto a cada um e a todos os objetos.
- Rodapé: Diversas informações que permitem o acesso rápido ao arquivo.

Nos interessa, neste exercício a informação `startxref` que é uma das últimas do arquivo. Você deve abrir o arquivo

```
arq06.pdf
```

publicado no lugar usual, com um editor de hexadecimal e deve procurar o startxref no final do arquivo. Na linha seguinte, existe um endereço que o PDF coloca em decimal. Você precisa transformar este número em hexadecimal e localizar no arquivo (mais acima) este endereço.

Nele, deve aparecer a palavra `xref` e na linha seguinte um zero e um número. Na linha seguinte, uma entrada (que todo arquivo tem) com muitos zeros, o número 65535 e uma letra. Na linha seguinte, **um número** seguido de zeros e uma letra. Trata-se do endereço do primeiro objeto do arquivo.

Responda aqui:

endereço da xref em hexadecimal	endereço do primeiro objeto em decimal

CEP - UP - UTFPR - PUCPr - UFPR –
17/06/2024 - 17:11:23.4
Prof Dr P Kantek (pkantek@gmail.com)
Objetos binários grandes vivxq10a, V: 1.01 76625
RAUL KNAPKI CUNHA 24CC2301 - 14 entregar ate 4/julho _____ / _____ / _____

Objetos binários grandes - BLOBs

O objetivo desta atividade e conhecer e manusear os chamados objetos binários grandes: imagens, vídeos, documentos, músicas, etc. Tais objetos quando estão na memória do computador são conhecidos como BLOBs e quando vão para uma memória não volátil (pendrives, discos, SSDs, nuvem, ...) passam a ser arquivos.

Resultados da digitalização crescente da nossa sociedade, precisam ser conhecidos para deles se extraírem todos os benefícios que eles apresentam.

Eis alguns raciocínios: a documentação em papel de um avião boeing 747 ocuparia 5 aviões 747 para ser transportada. Alguém ainda compra enciclopédias em papel ? Lembram de quando as fotografias eram analógicas ? Um filme Ektachrome (36 fotos) custava bem caro. Vou chutar uns 150 reais, o que daria quase 5 reais/foto, só o filme, sem contar a revelação e as ampliações. Quando o I.R. era em papel, a receita precisava montar uma operação de guerra, envolvendo todos os bancos, apenas para receber as declarações, e isso que era mais ou menos 1/3 do número de contribuintes atual. E as músicas: durante muitos anos, no dia do meu pagamento eu ia a uma loja de discos (?) e comprava 1 único LP, ($\pm$ 30 minutos de música) era o que o meu dinheiro alcançava. Aliás, falando de música, vale comentar a ascensão e queda do formato CD. Ele passou de maravilha tecnológica em meados dos 80 a algo obsoleto em meados dos 2010. Do céu ao inferno em 30 anos. E, se perder em uma cidade estranha, tendo apenas o guia 4 Rodas para ajudar ? Uma sensação indescritível.

Vamos estudar mais de perto alguns BLOBs, lembrando que qualquer um pode criar e chamar de seu, um BLOB. Pela sua definição (grande bloco de dados binários com uma lógica interna de organização e acesso) vê-se que isto é perfeitamente possível. Assim, vamos nos limitar a BLOBs padrão como formatos e conceitos conhecidos e publicados.

Imagem É difícil contar esta história. Ela começa no daguerrótipo de meados do século 19, através da fixação de imagens em placas de vidro. Segue pelas imagens de TV que eram capturadas e transmitidas eletronicamente. Seu registro foi feito em gravações magnéticas analógicas. Uma nova necessidade surgiu na obtenção de fotos na Lua e além, quando não havia como retornar filmes. A propósito, um pouco antes disto, a vigilância eletrônica na Guerra Fria era muito custosa: satélites fotografavam o inimigo e precisavam ejetar os filmes (que acabavam rápido) ao sobrevoar território amigo. A primeira foto da lua, transmitida eletronicamente foi feita pela espaçonave Ranger em 1964, 17 minutos antes dela se espatifar na superfície lunar. Quando a Internet começou (por volta de 1990) a necessidade de manusear imagens sofreu uma mudança: antes o gerador e o visualizador eram de mesma origem, pelo que não havia necessidade de padronização. Com a Internet essa história mudou. Uma das primeiras respostas a esta demanda foi o formato BMP (Bitmap Image File). Embora originalmente proposto pela Microsoft e IBM, teve seu padrão publicado e virou de fato, um formato livre. Sua origem é a bio-física dos olhos humanos (3 canais separados para Red, Green e Blue), além de um truque bem legal chamado mapeamento pelo qual se negocia tamanho do arquivo VERSUS menor quantidade de cores. A principal inovação do BMP foi criar, publicar e obedecer a um padrão. Parece pouco, mas não é. A chave da popularização e do sucesso sempre é um PADRãO. Anote isso na sua cabeça e nunca mais esqueça. O padrão pode ser ruim ou bom (neste caso não é muito), mas é um padrão. A imagem é uma matriz numérica (outra inovação entusiasmadora), logo sujeita a sofrer a ação maravilhosa de qualquer matemática.

Padrão BMP O arquivo começa com um bloco de controle de 54 bytes, que contém o identificador 'BM', o tamanho do objeto (arquivo), onde a imagem começa, L=largura e A=altura da imagem, profundidade). Depois, dependendo da profundidade, pode haver uma tabela de cores. Finalmente a imagem na forma de $L \times A$ se for uma imagem mapeada ou 3 tabelas de $L \times A$, uma para cada cor se for uma imagem true-color. Veja na imagem a seguir estes conceitos

Exemplo:

```
      0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
      -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
0000  42 4D 62 05 00 00 00 00 00 00 36 04 00 00 28 00 BM.....6...(
0010  00 00 12 00 00 00 0F 00 00 00 01 00 08 00 00 00 ................
0020  00 00 60 01 00 00 00 00 00 00 00 00 00 00 00 00 ..`.............
0030  00 00 00 00 00 00 00 00 FF FF 00 80 80 80 00 FF ................
0040  FF 00 FF 00 FF 00 00 00 00 FF 00 00 00 00 00 00 ................
0050  00 00 FF FF 00 00 FF 00 00 00 09 09 09 00 0A 0A ................
0060  0A 00 0B 0B 0B 00 0C 0C 0C 00 0D 0D 0D 00 0E 0E ................
0070  0E 00 0F 0F 0F 00 10 10 10 00 11 11 11 00 12 12 ................
...
0410  F6 00 F7 F7 F7 00 F8 F8 F8 00 F9 F9 F9 00 FA FA ................
0420  FA 00 FB FB FB 00 FC FC FC 00 FD FD FD 00 FE FE ................
0430  FE 00 FF FF FF 00 01 01 03 01 01 01 08 01 01 01 ................
0440  01 01 02 01 01 06 01 01 00 00 04 04 03 04 04 04 ................
0450  04 04 04 04 04 04 04 04 04 04 04 04 00 00 01 01 ................
```

Os problemas do BMP incluem: nenhuma compressão, o que graças ao fenômeno da *localidade de referência espacial* é uma tragédia. Também não prevê nenhum tipo de animação, o que faz falta na Internet. Não sabia (hoje sabe) lidar com transparência.

Um concorrente importante do BMP (lá nos primórdios) foi o formato GIF que admite compressão (usando o belo algoritmo LZW) e também admite uma forma rudimentar de animação. O problema do GIF é que ele foi e é um formato proprietário. Isto deu origem ao formato PNG que é praticamente o mesmo, mas livre.

A compressão poderia ser muito melhor, se se admitisse algum nivel de degradação na reconstituição da imagem após sua transformação em BLOB. Isto foi conseguido no padrão JPEG, que ao admitir degradações (sob controle), gera compressões muito maiores do que GIF ou similares. Outra vantagem é a possibilidade de criação em etapas das imagens (coisa que o BMP nunca ofereceu). O padrão foi formalizado pela ISO em 1991, mas na época disparou uma briga por patentes e direitos. Nos EUA, as últimas patentes expiraram em 2004. A compressão aqui é pelo uso da Transformada Discreta do Coseno. A sua operação está além do nosso interesse (querendo, olhe VIVX690c ou 690p), mas em resumo: a imagem é quebrada em blocos de 8×8; esta matriz é multiplicada pela matriz DCT, o que passa do domínio do pixel para o domínio da frequência (Transformada de Fourier); os coeficientes são quantizados; os coeficientes são comprimidos. Depois, se faz o inverso e *voilà*.

Som O MP3, sigla para MPEG-1 Audio Layer III, é um formato de arquivo de áudio digital que revolucionou a forma como consumimos e compartilhamos música.

Desenvolvido no início da década de 1990 pelo Moving Picture Experts Group (MPEG), o MP3 se tornou o formato dominante para música digital, utilizado em players de música portáteis, computadores, smartphones e serviços de streaming online. O MP3 utiliza uma técnica de compressão com perda chamada percepção auditiva psicoacústica para reduzir o tamanho dos arquivos de áudio. Essa técnica remove informações da faixa de áudio que o ouvido humano provavelmente não consegue perceber, resultando em arquivos menores sem comprometer significativamente a qualidade da música para a maioria dos ouvintes. Apesar da compressão, o MP3 oferece uma qualidade de som bastante boa, especialmente em taxas de bits mais altas (como 128 kbps ou superior). Tamanho de Arquivo Reduzido: Os arquivos MP3 são significativamente menores do que os arquivos de áudio não compactados, como CDs, permitindo fácil armazenamento e compartilhamento de música digital.

O MP3 teve um impacto profundo na indústria da música e na forma como consumimos música gerando um declínio das vendas de CDs: Os consumidores passaram a baixar e compartilhar músicas digitalmente. O MP3 impulsionou o crescimento da música digital, com o surgimento de serviços de streaming online como Spotify, Apple Music e Deezer. O MP3 abriu caminho para novos modelos de negócios na indústria da música, como a venda de músicas online e assinaturas de serviços de streaming.

Apesar de seu sucesso, o MP3 também enfrentou alguns desafios: Em taxas de bits mais baixas (como 64 kbps), a qualidade de som do MP3 pode ser perceptívelmente inferior à do áudio não compactado. Propriedades Intelectuais: Questões de direitos autorais e pirataria surgiram com a proliferação de arquivos MP3 compartilhados online. Formatos Alternativos: Novos formatos de áudio digital, como FLAC e AAC, oferecem maior qualidade de som ou taxas de compressão mais eficientes, mas ainda não alcançaram o mesmo nível de popularidade do MP3.

Outros padrões: MID (notação musical sintética), WAV (nenhuma perda de qualidade, logo arquivos grandes), além do formato OGG, da plataforma APPLE.

Vídeo No vídeo, estamos diante de necessidade maior de compressão. Se uma imagem estática já é grande, imagine precisar de 30 imagens estáticas por segundo. Agora, a localidade de referência ganha um novo viés, a l.r. no tempo. Ou seja, se imaginarmos um vídeo como um array 3-d (dimensões tempo, altura e largura), podemos comprimir em 2 dessas dimensões.

Aqui a história começa com os formatos analógicos (VHS, Betamax), segue pelo CD-ROM e VCD, e chega ao DVD. Depois vem os formatos AVC (padrão H.264), depois HEVC (H.265), VP9 (formato livre e aberto da Google) e AV1 (formato livre e aberto, melhor que o H.265). Diferentemente da imagem (e semelhantemente ao som), aqui existe muita preocupação com a latência (atraso no envio dos pacotes) que impacta diretamente a sincronia. O padrão H.264 utiliza timestamps e buffers para sincronizar áudio e vídeo. Em caso de perda de pacotes, mecanismos de recuperação de erros como o Concealment Motion Vector (CMV) podem ser utilizados para minimizar o impacto na sincronia. O CMV divide o vídeo em macroblocos. Quando há perda de um pacote contendo os vetores de movimento de um determinado macrobloco, o CMV busca os macroblocos vizinhos ($\uparrow, \downarrow, \leftarrow \rightarrow$) e ve se algum deles tem movimento similar ao esperado do macrobloco perdido. Daí ele "empresta" vetores e reconstitui o movimento ausente.

PDF O Portable Document Format (PDF), criado pela Adobe em 1993, se tornou um dos formatos de arquivo mais utilizados no mundo, presente em diversos setores e aplicações. Mais detalhes: vivx680a. Um documento PDF pode ser aberto e visualizado em qualquer dispositivo com um leitor de PDF instalado, independentemente do sistema operacional ou hardware utilizado. Isso garante que a formatação original do documento é sempre a mesma, independentemente da plataforma em que ele é aberto. Isso é crucial para documentos que exigem alta fidelidade visual, como apresentações, relatórios e publicações. O PDF oferece recursos de segurança robustos, como criptografia e assinaturas digitais, que protegem o conteúdo contra acessos não autorizados, modificações e falsificações. Isso torna o PDF ideal para documentos confidenciais e contratos legais. Até porque ele pode ser acompanhado de uma assinatura digital (*hash*). Suporta recursos de acessibilidade, como tags de estrutura e legendas para imagens, que facilitam o acesso ao conteúdo do documento por pessoas com deficiência visual ou outras necessidades especiais. Ao contrário de outros formatos de arquivo que podem ser corrompidos ou se tornar obsoletos com o tempo, o PDF é um formato durável e estável. Pense na economia de árvores: imagine se todos os PDFs existentes fossem para o papel. Ele é ideal para compartilhar documentos com outras pessoas, pois garante que a formatação original seja preservada e o conteúdo seja visualizado corretamente em diferentes dispositivos. O PDF é uma ótima opção para armazenar documentos importantes de forma segura e durável, pois protege o conteúdo contra modificações e garante que ele possa ser acessado no futuro. Revistas, livros, relatórios e outros materiais de publicação podem ser distribuídos no formato PDF, garantindo uma experiência de leitura consistente em diferentes dispositivos. **Outros:** Deixo de tratar aqui, por absoluta falta de espaço, outros BLOBs como exames médicos, arquivos de telescópios (vivx760), ...

Olhando dentro de um BLOB

Agora e hora de examinar com um "microscópio" as partes internas de um desses arquivos. O escolhido pela importância é o arquivo PDF. Pode parecer pouco, mas esta ideia simples, já salvou centenas de milhares (milhões ?) de árvores. Já são muitos os sistemas de informação que não geram mais saídas em papel, limitando-se a criar arquivos PDF.

Um arquivo PDF descreve como uma página ficará depois de impressa e ele é melhor do que imagem dessa página pelas seguintes razões

- ⋆ A imagem como tal sempre é muito maior do que sua descrição. A razão para isso é a evidente redundância de qualquer imagem.

⋆ A imagem como tal é fixa e só pode ser rotacionada, redimensionada ou de alguma maneira modificada à custa de processamento e perda de qualidade.

⋆ Sobre uma imagem não há como manipular o conteúdo, por exemplo conduzindo uma busca textual (^F alguma coisa ou então ^C e ^V).

⋆ Alterar uma única palavra sobre uma imagem exige habilidades de pintura. Sobre a descrição de uma página é tarefa simples: basta trocar uma palavra pela outra.

⋆ Imagine uma imagem enviada para um monitor (72 DPP) ou a mesma imagem enviada para uma impressora (300 DPP). Não podem ser a mesma imagem.

A versão 1.0 do padrão PDF nasceu em 1993 quando a Adobe lançou 2 programas: o Distiller (para gerar PDF) e o Reader (para ler): ambos pagos. A coisa começou a mudar quando o departamento de rendas americano comprou uma licença que permitia aos contribuintes baixar o Reader. Rapidamente a Adobe deu-se conta de que seria melhor para ela distribuir gratuitamente o Reader. Nos dez anos seguintes o PDF acabou se tornando o padrão de fato de descrição de imagens.

Apresenta as seguintes vantagens:

⋆ acesso aleatório: qualquer página pode ser montada independente das demais.

⋆ linearização: todos os elementos necessários para montar uma página estão juntos.

⋆ atualização incremental: mudanças ficam registradas no fim do arquivo. Vantagem: opção de desfazer ilimitada e rápida atualização. Desvantagem: o arquivo não pára de crescer.

⋆ fontes incorporadas: o destino sempre será montado corretamente. A fonte é desbastada de tudo que não é necessário.

⋆ padronização ISO: em 2008 a ISO abraçou o padrão Adobe PDF 1.7, o que o transformou em padrão aberto.

⋆ compatibilidade para trás e para a frente: Qualquer leitor lê as versões antigas e deve ler as futuras, já que todos os leitores ignoram o que não conhecem.

⋆ integrado a todos os principais browsers do protocolo `HTTP`.

Um arquivo PDF

Cabeçalho Formado por 2 linhas: a primeira identifica o arquivo como um PDF e informa a versão PDF em que ele foi gerado. A segunda linha inclui caracteres que não podem ser impressos (para avisar eventuais leitores deste fato). Exemplo:

```
%PDF-1.0
%âã
```

Corpo Um conjunto de nodos de um grafo, contendo as páginas, conteúdo gráfico, textos, sempre na forma de objetos. Cada objeto começa com um número de objeto (iniciando em 1), um número de geração (sempre 0) e a palavra chave `obj`. O objeto termina pela palavra chave `endobj`. Exemplo:

```
1 0 obj     -- descrevendo o objeto 1
<<          -- começa um dicionário
 /Kids [2 0 R]    -- /Kids tem o valor 2 0 R
 /Count 1         -- só um
 /Type /Pages     -- do tipo : página
>>               -- fim do dicionário
endobj            -- fim do objeto
```

Tabela de Referência Cruzada Lista o deslocamento em bytes de cada objeto no corpo do arquivo. Exemplo

```
0 6          -- seis entradas na tabela
0000000000  65535  f   -- entrada especial
0000000015  00000  n   -- obj1 no desloc 15
0000000074  00000  n   -- obj2 no desloc 74
...
```

Rodapé Informações e metadados para aumentar a performance de acesso. A primeira entrada é `trailer`. Depois o dicionário de trailer que deve apresentar ao menos a entrada `\Size` dizendo quantos itens há na tabela de referência cruzada e `\Root` dizendo qual objeto é o catálogo do documento. Depois vem uma linha que só contém `startxref` seguida por uma única linha indicando o deslocamento em bytes utilizado no inicio da tabela de referência cruzada do arquivo. Termina tudo a linha `%%EOF`. Exemplo:

```
trailer   -- palavra chave
<<   -- começa um dicionário
/Root 5 0 R   -- o catálogo é o objeto 5
/Size 6       -- tem o tamanho 6
>>   -- fim do dicionário
startxref
459  -- deslocamento da tabela de ref cruzada
%%EOF   -- fim de arquivo
```

Componentes

Uma descrição PDF suporta 5 componentes básicos: i. Números inteiros e reais (mas não exponenciais); ii. strings que são escritos entre parênteses; iii.Nomes, que sempre começam por uma barra normal; iv.Os valores booleanos `true` e `false` e v. O componente nulo (`null`). Aceita também 3 componentes compostos: i. Arrays, que são uma coleção ordenada de outros componentes, escritos entre colchetes; ii.Dicionários que são uma lista não ordenada de duplas: nome e componente. Começam por `<<` e terminam por `>>`. Finalmente iii. Fluxos, que armazenam dados binários, sempre acompanhados de um dicionário que descreve seus atributos, como comprimento e parâmetros de compactação, por exemplo.

☞ Para você fazer

As coisas em um arquivo PDF são separadas por um divisor de linhas. Pode ser o caracter $X'10'$, ou o $X'13'$ ou uma combinação de ambos. Ao examinar um PDF em hexadecimal, acostume-se a separar as linhas marcando este caractere. Isto posto, os componentes principais do arquivo PDF são:

- Header: identificação PDF e uma coleção de letras acentuadas
- Corpo: objetos numerados a partir de 1 sequencialmente, e identificados pela palavra `obj` no início e `endobj` no final. Esses objetos podem ser textos, desenhos, fontes, imagens, metadados, links, e um monte de coisas mais.
- Tabela de referência: uma lista de onde (no arquivo) está cada objeto. Esta tabela permite o acesso direto a cada um e a todos os objetos.
- Rodapé: Diversas informações que permitem o acesso rápido ao arquivo.

Nos interessa, neste exercício a informação `startxref` que é uma das últimas do arquivo. Você deve abrir o arquivo

```
arq02.pdf
```

publicado no lugar usual, com um editor de hexadecimal e deve procurar o startxref no final do arquivo. Na linha seguinte, existe um endereço que o PDF coloca em decimal. Você precisa transformar este número em hexadecimal e localizar no arquivo (mais acima) este endereço.

Nele, deve aparecer a palavra `xref` e na linha seguinte um zero e um número. Na linha seguinte, uma entrada (que todo arquivo tem) com muitos zeros, o número 65535 e uma letra. Na linha seguinte, **um número** seguido de zeros e uma letra. Trata-se do endereço do primeiro objeto do arquivo.

Responda aqui:

endereço da xref em hexadecimal	endereço do primeiro objeto em decimal

CEP - UP - UTFPR - PUCPr - UFPR –
17/06/2024 - 17:11:23.4
Prof Dr P Kantek (pkantek@gmail.com)
Objetos binários grandes vivxq10a, V: 1.01 76632
VICTOR BRONDANI DOS SANTOS 24CC2301
- 15 entregar ate 4/julho _____ / _____ / _____

Objetos binários grandes - BLOBs

O objetivo desta atividade e conhecer e manusear os chamados objetos binários grandes: imagens, vídeos, documentos, músicas, etc. Tais objetos quando estão na memória do computador são conhecidos como BLOBs e quando vão para uma memória não volátil (pendrives, discos, SSDs, nuvem, ...) passam a ser arquivos.

Resultados da digitalização crescente da nossa sociedade, precisam ser conhecidos para deles se extraírem todos os benefícios que eles apresentam.

Eis alguns raciocínios: a documentação em papel de um avião boeing 747 ocuparia 5 aviões 747 para ser transportada. Alguém ainda compra enciclopédias em papel ? Lembram de quando as fotografias eram analógicas ? Um filme Ektachrome (36 fotos) custava bem caro. Vou chutar uns 150 reais, o que daria quase 5 reais/foto, só o filme, sem contar a revelação e as ampliações. Quando o I.R. era em papel, a receita precisava montar uma operação de guerra, envolvendo todos os bancos, apenas para receber as declarações, e isso que era mais ou menos 1/3 do número de contribuintes atual. E as músicas: durante muitos anos, no dia do meu pagamento eu ia a uma loja de discos (?) e comprava 1 único LP, ($\pm$ 30 minutos de música) era o que o meu dinheiro alcançava. Aliás, falando de música, vale comentar a ascensão e queda do formato CD. Ele passou de maravilha tecnológica em meados dos 80 a algo obsoleto em meados dos 2010. Do céu ao inferno em 30 anos. E, se perder em uma cidade estranha, tendo apenas o guia 4 Rodas para ajudar ? Uma sensação indescritível.

Vamos estudar mais de perto alguns BLOBs, lembrando que qualquer um pode criar e chamar de seu, um BLOB. Pela sua definição (grande bloco de dados binários com uma lógica interna de organização e acesso) vê-se que isto é perfeitamente possível. Assim, vamos nos limitar a BLOBs padrão como formatos e conceitos conhecidos e publicados.

Imagem É difícil contar esta história. Ela começa no daguerrótipo de meados do século 19, através da fixação de imagens em placas de vidro. Segue pelas imagens de TV que eram capturadas e transmitidas eletronicamente. Seu registro foi feito em gravações magnéticas analógicas. Uma nova necessidade surgiu na obtenção de fotos na Lua e além, quando não havia como retornar filmes. A propósito, um pouco antes disto, a vigilância eletrônica na Guerra Fria era muito custosa: satélites fotografavam o inimigo e precisavam ejetar os filmes (que acabavam rápido) ao sobrevoar território amigo. A primeira foto da lua, transmitida eletronicamente foi feita pela espaçonave Ranger em 1964, 17 minutos antes dela se espatifar na superfície lunar. Quando a Internet começou (por volta de 1990) a necessidade de manusear imagens sofreu uma mudança: antes o gerador e o visualizador eram de mesma origem, pelo que não havia necessidade de padronização. Com a Internet essa história mudou. Uma das primeiras respostas a esta demanda foi o formato BMP (Bitmap Image File). Embora originalmente proposto pela Microsoft e IBM, teve seu padrão publicado e virou de fato, um formato livre. Sua origem é a bio-física dos olhos humanos (3 canais separados para Red, Green e Blue), além de um truque bem legal chamado mapeamento pelo qual se negocia tamanho do arquivo VERSUS menor quantidade de cores. A principal inovação do BMP foi criar, publicar e obedecer a um padrão. Parece pouco, mas não é. A chave da popularização e do sucesso sempre é um PADRãO. Anote isso na sua cabeça e nunca mais esqueça. O padrão pode ser ruim ou bom (neste caso não é muito), mas é um padrão. A imagem é uma matriz numérica (outra inovação entusiasmadora), logo sujeita a sofrer a ação maravilhosa de qualquer matemática.

Padrão BMP O arquivo começa com um bloco de controle de 54 bytes, que contém o identificador 'BM', o tamanho do objeto (arquivo), onde a imagem começa, L=largura e A=altura da imagem, profundidade). Depois, dependendo da profundidade, pode haver uma tabela de cores. Finalmente a imagem na forma de $L \times A$ se for uma imagem mapeada ou 3 tabelas de $L \times A$, uma para cada cor se for uma imagem true-color. Veja na imagem a seguir estes conceitos

Exemplo:

```
      0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
      -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
0000  42 4D 62 05 00 00 00 00 00 00 36 04 00 00 28 00 BM.....6...(
0010  00 00 12 00 00 00 0F 00 00 00 01 00 08 00 00 00 ................
0020  00 00 60 01 00 00 00 00 00 00 00 00 00 00 00 00 ..`.............
0030  00 00 00 00 00 00 00 00 FF FF 00 80 80 80 00 FF ................
0040  FF 00 FF 00 FF 00 00 00 00 FF 00 00 00 00 00 00 ................
0050  00 00 FF FF 00 00 FF 00 00 00 09 09 09 00 0A 0A ................
0060  0A 00 0B 0B 0B 00 0C 0C 0C 00 0D 0D 0D 00 0E 0E ................
0070  0E 00 0F 0F 0F 00 10 10 10 00 11 11 11 00 12 12 ................
...
0410  F6 00 F7 F7 F7 00 F8 F8 F8 00 F9 F9 F9 00 FA FA ................
0420  FA 00 FB FB FB 00 FC FC FC 00 FD FD FD 00 FE FE ................
0430  FE 00 FF FF FF 00 01 01 03 01 01 01 08 01 01 01 ................
0440  01 01 02 01 01 06 01 01 00 00 04 04 03 04 04 04 ................
0450  04 04 04 04 04 04 04 04 04 04 04 04 00 00 01 01 ................
```

Os problemas do BMP incluem: nenhuma compressão, o que graças ao fenômeno da *localidade de referência espacial* é uma tragédia. Também não prevê nenhum tipo de animação, o que faz falta na Internet. Não sabia (hoje sabe) lidar com transparência.

Um concorrente importante do BMP (lá nos primórdios) foi o formato GIF que admite compressão (usando o belo algoritmo LZW) e também admite uma forma rudimentar de animação. O problema do GIF é que ele foi e é um formato proprietário. Isto deu origem ao formato PNG que é praticamente o mesmo, mas livre.

A compressão poderia ser muito melhor, se se admitisse algum nivel de degradação na reconstituição da imagem após sua transformação em BLOB. Isto foi conseguido no padrão JPEG, que ao admitir degradações (sob controle), gera compressões muito maiores do que GIF ou similares. Outra vantagem é a possibilidade de criação em etapas das imagens (coisa que o BMP nunca ofereceu). O padrão foi formalizado pela ISO em 1991, mas na época disparou uma briga por patentes e direitos. Nos EUA, as últimas patentes expiraram em 2004. A compressão aqui é pelo uso da Transformada Discreta do Coseno. A sua operação está além do nosso interesse (querendo, olhe VIVX690c ou 690p), mas em resumo: a imagem é quebrada em blocos de 8×8; esta matriz é multiplicada pela matriz DCT, o que passa do domínio do pixel para o domínio da frequência (Transformada de Fourier); os coeficientes são quantizados; os coeficientes são comprimidos. Depois, se faz o inverso e *voilà*.

Som O MP3, sigla para MPEG-1 Audio Layer III, é um formato de arquivo de áudio digital que revolucionou a forma como consumimos e compartilhamos música.

Desenvolvido no início da década de 1990 pelo Moving Picture Experts Group (MPEG), o MP3 se tornou o formato dominante para música digital, utilizado em players de música portáteis, computadores, smartphones e serviços de streaming online. O MP3 utiliza uma técnica de compressão com perda chamada percepção auditiva psicoacústica para reduzir o tamanho dos arquivos de áudio. Essa técnica remove informações da faixa de áudio que o ouvido humano provavelmente não consegue perceber, resultando em arquivos menores sem comprometer significativamente a qualidade da música para a maioria dos ouvintes. Apesar da compressão, o MP3 oferece uma qualidade de som bastante boa, especialmente em taxas de bits mais altas (como 128 kbps ou superior). Tamanho de Arquivo Reduzido: Os arquivos MP3 são significativamente menores do que os arquivos de áudio não compactados, como CDs, permitindo fácil armazenamento e compartilhamento de música digital.

O MP3 teve um impacto profundo na indústria da música e na forma como consumimos música gerando um declínio das vendas de CDs: Os consumidores passaram a baixar e compartilhar músicas digitalmente. O MP3 impulsionou o crescimento da música digital, com o surgimento de serviços de streaming online como Spotify, Apple Music e Deezer. O MP3 abriu caminho para novos modelos de negócios na indústria da música, como a venda de músicas online e assinaturas de serviços de streaming.

Apesar de seu sucesso, o MP3 também enfrentou alguns desafios: Em taxas de bits mais baixas (como 64 kbps), a qualidade de som do MP3 pode ser perceptívelmente inferior à do áudio não compactado. Propriedades Intelectuais: Questões de direitos autorais e pirataria surgiram com a proliferação de arquivos MP3 compartilhados online. Formatos Alternativos: Novos formatos de áudio digital, como FLAC e AAC, oferecem maior qualidade de som ou taxas de compressão mais eficientes, mas ainda não alcançaram o mesmo nível de popularidade do MP3.

Outros padrões: MID (notação musical sintética), WAV (nenhuma perda de qualidade, logo arquivos grandes), além do formato OGG, da plataforma APPLE.

Vídeo No vídeo, estamos diante de necessidade maior de compressão. Se uma imagem estática já é grande, imagine precisar de 30 imagens estáticas por segundo. Agora, a localidade de referência ganha um novo viés, a l.r. no tempo. Ou seja, se imaginarmos um vídeo como um array 3-d (dimensões tempo, altura e largura), podemos comprimir em 2 dessas dimensões.

Aqui a história começa com os formatos analógicos (VHS, Betamax), segue pelo CD-ROM e VCD, e chega ao DVD. Depois vem os formatos AVC (padrão H.264), depois HEVC (H.265), VP9 (formato livre e aberto da Google) e AV1 (formato livre e aberto, melhor que o H.265). Diferentemente da imagem (e semelhantemente ao som), aqui existe muita preocupação com a latência (atraso no envio dos pacotes) que impacta diretamente a sincronia. O padrão H.264 utiliza timestamps e buffers para sincronizar áudio e vídeo. Em caso de perda de pacotes, mecanismos de recuperação de erros como o Concealment Motion Vector (CMV) podem ser utilizados para minimizar o impacto na sincronia. O CMV divide o vídeo em macroblocos. Quando há perda de um pacote contendo os vetores de movimento de um determinado macrobloco, o CMV busca os macroblocos vizinhos ($\uparrow, \downarrow, \leftarrow \rightarrow$) e ve se algum deles tem movimento similar ao esperado do macrobloco perdido. Daí ele "empresta" vetores e reconstitui o movimento ausente.

PDF O Portable Document Format (PDF), criado pela Adobe em 1993, se tornou um dos formatos de arquivo mais utilizados no mundo, presente em diversos setores e aplicações. Mais detalhes: vivx680a. Um documento PDF pode ser aberto e visualizado em qualquer dispositivo com um leitor de PDF instalado, independentemente do sistema operacional ou hardware utilizado. Isso garante que a formatação original do documento é sempre a mesma, independentemente da plataforma em que ele é aberto. Isso é crucial para documentos que exigem alta fidelidade visual, como apresentações, relatórios e publicações. O PDF oferece recursos de segurança robustos, como criptografia e assinaturas digitais, que protegem o conteúdo contra acessos não autorizados, modificações e falsificações. Isso torna o PDF ideal para documentos confidenciais e contratos legais. Até porque ele pode ser acompanhado de uma assinatura digital (*hash*). Suporta recursos de acessibilidade, como tags de estrutura e legendas para imagens, que facilitam o acesso ao conteúdo do documento por pessoas com deficiência visual ou outras necessidades especiais. Ao contrário de outros formatos de arquivo que podem ser corrompidos ou se tornar obsoletos com o tempo, o PDF é um formato durável e estável. Pense na economia de árvores: imagine se todos os PDFs existentes fossem para o papel. Ele é ideal para compartilhar documentos com outras pessoas, pois garante que a formatação original seja preservada e o conteúdo seja visualizado corretamente em diferentes dispositivos. O PDF é uma ótima opção para armazenar documentos importantes de forma segura e durável, pois protege o conteúdo contra modificações e garante que ele possa ser acessado no futuro. Revistas, livros, relatórios e outros materiais de publicação podem ser distribuídos no formato PDF, garantindo uma experiência de leitura consistente em diferentes dispositivos. **Outros:** Deixo de tratar aqui, por absoluta falta de espaço, outros BLOBs como exames médicos, arquivos de telescópios (vivx760), ...

Olhando dentro de um BLOB

Agora e hora de examinar com um "microscópio" as partes internas de um desses arquivos. O escolhido pela importância é o arquivo PDF. Pode parecer pouco, mas esta ideia simples, já salvou centenas de milhares (milhões ?) de árvores. Já são muitos os sistemas de informação que não geram mais saídas em papel, limitando-se a criar arquivos PDF.

Um arquivo PDF descreve como uma página ficará depois de impressa e ele é melhor do que imagem dessa página pelas seguintes razões

- ⋆ A imagem como tal sempre é muito maior do que sua descrição. A razão para isso é a evidente redundância de qualquer imagem.

⋆ A imagem como tal é fixa e só pode ser rotacionada, redimensionada ou de alguma maneira modificada à custa de processamento e perda de qualidade.

⋆ Sobre uma imagem não há como manipular o conteúdo, por exemplo conduzindo uma busca textual (^F alguma coisa ou então ^C e ^V).

⋆ Alterar uma única palavra sobre uma imagem exige habilidades de pintura. Sobre a descrição de uma página é tarefa simples: basta trocar uma palavra pela outra.

⋆ Imagine uma imagem enviada para um monitor (72 DPP) ou a mesma imagem enviada para uma impressora (300 DPP). Não podem ser a mesma imagem.

A versão 1.0 do padrão PDF nasceu em 1993 quando a Adobe lançou 2 programas: o Distiller (para gerar PDF) e o Reader (para ler): ambos pagos. A coisa começou a mudar quando o departamento de rendas americano comprou uma licença que permitia aos contribuintes baixar o Reader. Rapidamente a Adobe deu-se conta de que seria melhor para ela distribuir gratuitamente o Reader. Nos dez anos seguintes o PDF acabou se tornando o padrão de fato de descrição de imagens.

Apresenta as seguintes vantagens:

⋆ acesso aleatório: qualquer página pode ser montada independente das demais.

⋆ linearização: todos os elementos necessários para montar uma página estão juntos.

⋆ atualização incremental: mudanças ficam registradas no fim do arquivo. Vantagem: opção de desfazer ilimitada e rápida atualização. Desvantagem: o arquivo não pára de crescer.

⋆ fontes incorporadas: o destino sempre será montado corretamente. A fonte é desbastada de tudo que não é necessário.

⋆ padronização ISO: em 2008 a ISO abraçou o padrão Adobe PDF 1.7, o que o transformou em padrão aberto.

⋆ compatibilidade para trás e para a frente: Qualquer leitor lê as versões antigas e deve ler as futuras, já que todos os leitores ignoram o que não conhecem.

⋆ integrado a todos os principais browsers do protocolo `HTTP`.

Um arquivo PDF

Cabeçalho Formado por 2 linhas: a primeira identifica o arquivo como um PDF e informa a versão PDF em que ele foi gerado. A segunda linha inclui caracteres que não podem ser impressos (para avisar eventuais leitores deste fato). Exemplo:

```
%PDF-1.0
%âã
```

Corpo Um conjunto de nodos de um grafo, contendo as páginas, conteúdo gráfico, textos, sempre na forma de objetos. Cada objeto começa com um número de objeto (iniciando em 1), um número de geração (sempre 0) e a palavra chave `obj`. O objeto termina pela palavra chave `endobj`. Exemplo:

```
1 0 obj     -- descrevendo o objeto 1
<<          -- começa um dicionário
 /Kids [2 0 R]    -- /Kids tem o valor 2 0 R
 /Count 1         -- só um
 /Type /Pages     -- do tipo : página
>>               -- fim do dicionário
endobj            -- fim do objeto
```

Tabela de Referência Cruzada Lista o deslocamento em bytes de cada objeto no corpo do arquivo. Exemplo

```
0 6          -- seis entradas na tabela
0000000000  65535  f   -- entrada especial
0000000015  00000  n   -- obj1 no desloc 15
0000000074  00000  n   -- obj2 no desloc 74
...
```

Rodapé Informações e metadados para aumentar a performance de acesso. A primeira entrada é `trailer`. Depois o dicionário de trailer que deve apresentar ao menos a entrada `\Size` dizendo quantos itens há na tabela de referência cruzada e `\Root` dizendo qual objeto é o catálogo do documento. Depois vem uma linha que só contém `startxref` seguida por uma única linha indicando o deslocamento em bytes utilizado no inicio da tabela de referência cruzada do arquivo. Termina tudo a linha `%%EOF`. Exemplo:

```
trailer   -- palavra chave
<<   -- começa um dicionário
/Root 5 0 R   -- o catálogo é o objeto 5
/Size 6       -- tem o tamanho 6
>>   -- fim do dicionário
startxref
459  -- deslocamento da tabela de ref cruzada
%%EOF   -- fim de arquivo
```

Componentes

Uma descrição PDF suporta 5 componentes básicos: i. Números inteiros e reais (mas não exponenciais); ii. strings que são escritos entre parênteses; iii.Nomes, que sempre começam por uma barra normal; iv.Os valores booleanos `true` e `false` e v. O componente nulo (`null`). Aceita também 3 componentes compostos: i. Arrays, que são uma coleção ordenada de outros componentes, escritos entre colchetes; ii.Dicionários que são uma lista não ordenada de duplas: nome e componente. Começam por `<<` e terminam por `>>`. Finalmente iii. Fluxos, que armazenam dados binários, sempre acompanhados de um dicionário que descreve seus atributos, como comprimento e parâmetros de compactação, por exemplo.

☞ Para você fazer

As coisas em um arquivo PDF são separadas por um divisor de linhas. Pode ser o caracter $X'10'$, ou o $X'13'$ ou uma combinação de ambos. Ao examinar um PDF em hexadecimal, acostume-se a separar as linhas marcando este caractere. Isto posto, os componentes principais do arquivo PDF são:

- Header: identificação PDF e uma coleção de letras acentuadas
- Corpo: objetos numerados a partir de 1 sequencialmente, e identificados pela palavra `obj` no início e `endobj` no final. Esses objetos podem ser textos, desenhos, fontes, imagens, metadados, links, e um monte de coisas mais.
- Tabela de referência: uma lista de onde (no arquivo) está cada objeto. Esta tabela permite o acesso direto a cada um e a todos os objetos.
- Rodapé: Diversas informações que permitem o acesso rápido ao arquivo.

Nos interessa, neste exercício a informação `startxref` que é uma das últimas do arquivo. Você deve abrir o arquivo

```
arq08.pdf
```

publicado no lugar usual, com um editor de hexadecimal e deve procurar o startxref no final do arquivo. Na linha seguinte, existe um endereço que o PDF coloca em decimal. Você precisa transformar este número em hexadecimal e localizar no arquivo (mais acima) este endereço.

Nele, deve aparecer a palavra `xref` e na linha seguinte um zero e um número. Na linha seguinte, uma entrada (que todo arquivo tem) com muitos zeros, o número 65535 e uma letra. Na linha seguinte, **um número** seguido de zeros e uma letra. Trata-se do endereço do primeiro objeto do arquivo.

Responda aqui:

endereço da xref em hexadecimal	endereço do primeiro objeto em decimal

CEP - UP - UTFPR - PUCPr - UFPR –
17/06/2024 - 17:11:23.4
Prof Dr P Kantek (pkantek@gmail.com)
Objetos binários grandes vivxq10a, V: 1.01 76649
VICTOR HUGO DOS SANTOS DE CAMA
24CC2301 - 16 entregar ate 4/julho _____ /
/ _____

Objetos binários grandes - BLOBs

O objetivo desta atividade e conhecer e manusear os chamados objetos binários grandes: imagens, vídeos, documentos, músicas, etc. Tais objetos quando estão na memória do computador são conhecidos como BLOBs e quando vão para uma memória não volátil (pendrives, discos, SSDs, nuvem, ...) passam a ser arquivos.

Resultados da digitalização crescente da nossa sociedade, precisam ser conhecidos para deles se extraírem todos os benefícios que eles apresentam.

Eis alguns raciocínios: a documentação em papel de um avião boeing 747 ocuparia 5 aviões 747 para ser transportada. Alguém ainda compra enciclopédias em papel ? Lembram de quando as fotografias eram analógicas ? Um filme Ektachrome (36 fotos) custava bem caro. Vou chutar uns 150 reais, o que daria quase 5 reais/foto, só o filme, sem contar a revelação e as ampliações. Quando o I.R. era em papel, a receita precisava montar uma operação de guerra, envolvendo todos os bancos, apenas para receber as declarações, e isso que era mais ou menos 1/3 do número de contribuintes atual. E as músicas: durante muitos anos, no dia do meu pagamento eu ia a uma loja de discos (?) e comprava 1 único LP, ($\pm$ 30 minutos de música) era o que o meu dinheiro alcançava. Aliás, falando de música, vale comentar a ascensão e queda do formato CD. Ele passou de maravilha tecnológica em meados dos 80 a algo obsoleto em meados dos 2010. Do céu ao inferno em 30 anos. E, se perder em uma cidade estranha, tendo apenas o guia 4 Rodas para ajudar ? Uma sensação indescritível.

Vamos estudar mais de perto alguns BLOBs, lembrando que qualquer um pode criar e chamar de seu, um BLOB. Pela sua definição (grande bloco de dados binários com uma lógica interna de organização e acesso) vê-se que isto é perfeitamente possível. Assim, vamos nos limitar a BLOBs padrão como formatos e conceitos conhecidos e publicados.

Imagem É difícil contar esta história. Ela começa no daguerrótipo de meados do século 19, através da fixação de imagens em placas de vidro. Segue pelas imagens de TV que eram capturadas e transmitidas eletronicamente. Seu registro foi feito em gravações magnéticas analógicas. Uma nova necessidade surgiu na obtenção de fotos na Lua e além, quando não havia como retornar filmes. A propósito, um pouco antes disto, a vigilância eletrônica na Guerra Fria era muito custosa: satélites fotografavam o inimigo e precisavam ejetar os filmes (que acabavam rápido) ao sobrevoar território amigo. A primeira foto da lua, transmitida eletronicamente foi feita pela espaçonave Ranger em 1964, 17 minutos antes dela se espatifar na superfície lunar. Quando a Internet começou (por volta de 1990) a necessidade de manusear imagens sofreu uma mudança: antes o gerador e o visualizador eram de mesma origem, pelo que não havia necessidade de padronização. Com a Internet essa história mudou. Uma das primeiras respostas a esta demanda foi o formato BMP (Bitmap Image File). Embora originalmente proposto pela Microsoft e IBM, teve seu padrão publicado e virou de fato, um formato livre. Sua origem é a bio-física dos olhos humanos (3 canais separados para Red, Green e Blue), além de um truque bem legal chamado mapeamento pelo qual se negocia tamanho do arquivo VERSUS menor quantidade de cores. A principal inovação do BMP foi criar, publicar e obedecer a um padrão. Parece pouco, mas não é. A chave da popularização e do sucesso sempre é um PADRãO. Anote isso na sua cabeça e nunca mais esqueça. O padrão pode ser ruim ou bom (neste caso não é muito), mas é um padrão. A imagem é uma matriz numérica (outra inovação entusiasmadora), logo sujeita a sofrer a ação maravilhosa de qualquer matemática.

Padrão BMP O arquivo começa com um bloco de controle de 54 bytes, que contém o identificador 'BM', o tamanho do objeto (arquivo), onde a imagem começa, L=largura e A=altura da imagem, profundidade). Depois, dependendo da profundidade, pode haver uma tabela de cores. Finalmente a imagem na forma de $L \times A$ se for uma imagem mapeada ou 3 tabelas de $L \times A$, uma para cada cor se for uma imagem true-color. Veja na imagem a seguir estes conceitos

```
Exemplo:
        0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
       -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
0000   42 4D 62 05 00 00 00 00 00 00 36 04 00 00 28 00 BM.....6...(
0010   00 00 12 00 00 00 0F 00 00 00 01 00 08 00 00 00 ................
0020   00 00 60 01 00 00 00 00 00 00 00 00 00 00 00 01 ................
0030   00 00 00 00 00 00 00 00 FF FF 00 80 80 80 00 FF FF ............
0040   FF 00 FF 00 FF 00 00 00 FF 00 00 00 00 00 00 FF ................
0050   00 00 FF FF 00 00 FF 00 00 00 09 09 09 00 0A 0A ................
0060   0A 00 0B 0B 0B 00 0C 0C 0C 00 0D 0D 0D 00 0E 0E ................
0070   0E 00 0F 0F 0F 00 10 10 10 00 11 11 11 00 12 12 ................
...
0410   F6 00 F7 F7 F7 00 F8 F8 F8 00 F9 F9 F9 00 FA FA ................
0420   FA 00 FB FB FB 00 FC FC FC 00 FD FD FD 00 FE FE ................
0430   FE 00 FF FF FF 00 01 01 03 01 01 01 08 01 01 01 ................
0440   01 01 02 01 01 06 01 01 00 00 04 04 03 04 04 04 ................
0450   04 04 04 04 04 04 04 04 04 04 04 04 00 00 01 01 ................
```

Os problemas do BMP incluem: nenhuma compressão, o que graças ao fenômeno da *localidade de referência espacial* é uma tragédia. Também não prevê nenhum tipo de animação, o que faz falta na Internet. Não sabia (hoje sabe) lidar com transparência.

Um concorrente importante do BMP (lá nos primórdios) foi o formato GIF que admite compressão (usando o belo algoritmo LZW) e também admite uma forma rudimentar de animação. O problema do GIF é que ele foi e é um formato proprietário. Isto deu origem ao formato PNG que é praticamente o mesmo, mas livre.

A compressão poderia ser muito melhor, se se admitisse algum nivel de degradação na reconstituição da imagem após sua transformação em BLOB. Isto foi conseguido no padrão JPEG, que ao admitir degradações (sob controle), gera compressões muito maiores do que GIF ou similares. Outra vantagem é a possibilidade de criação em etapas das imagens (coisa que o BMP nunca ofereceu). O padrão foi formalizado pela ISO em 1991, mas na época disparou uma briga por patentes e direitos. Nos EUA, as últimas patentes expiraram em 2004. A compressão aqui é pelo uso da Transformada Discreta do Coseno. A sua operação está além do nosso interesse (querendo, olhe VIVX690c ou 690p), mas em resumo: a imagem é quebrada em blocos de 8×8; esta matriz é multiplicada pela matriz DCT, o que passa do domínio do pixel para o domínio da frequência (Transformada de Fourier); os coeficientes são quantizados; os coeficientes são comprimidos. Depois, se faz o inverso e *voilà*.

Som O MP3, sigla para MPEG-1 Audio Layer III, é um formato de arquivo de áudio digital que revolucionou a forma como consumimos e compartilhamos música.

Desenvolvido no início da década de 1990 pelo Moving Picture Experts Group (MPEG), o MP3 se tornou o formato dominante para música digital, utilizado em players de música portáteis, computadores, smartphones e serviços de streaming online. O MP3 utiliza uma técnica de compressão com perda chamada percepção auditiva psicoacústica para reduzir o tamanho dos arquivos de áudio. Essa técnica remove informações da faixa de áudio que o ouvido humano provavelmente não consegue perceber, resultando em arquivos menores sem comprometer significativamente a qualidade da música para a maioria dos ouvintes. Apesar da compressão, o MP3 oferece uma qualidade de som bastante boa, especialmente em taxas de bits mais altas (como 128 kbps ou superior). Tamanho de Arquivo Reduzido: Os arquivos MP3 são significativamente menores do que os arquivos de áudio não compactados, como CDs, permitindo fácil armazenamento e compartilhamento de música digital.

O MP3 teve um impacto profundo na indústria da música e na forma como consumimos música gerando um declínio das vendas de CDs: Os consumidores passaram a baixar e compartilhar músicas digitalmente. O MP3 impulsionou o crescimento da música digital, com o surgimento de serviços de streaming online como Spotify, Apple Music e Deezer. O MP3 abriu caminho para novos modelos de negócios na indústria da música, como a venda de músicas online e assinaturas de serviços de streaming.

Apesar de seu sucesso, o MP3 também enfrentou alguns desafios: Em taxas de bits mais baixas (como 64 kbps), a qualidade de som do MP3 pode ser perceptívelmente inferior à do áudio não compactado. Propriedades Intelectuais: Questões de direitos autorais e pirataria surgiram com a proliferação de arquivos MP3 compartilhados online. Formatos Alternativos: Novos formatos de áudio digital, como FLAC e AAC, oferecem maior qualidade de som ou taxas de compressão mais eficientes, mas ainda não alcançaram o mesmo nível de popularidade do MP3.

Outros padrões: MID (notação musical sintética), WAV (nenhuma perda de qualidade, logo arquivos grandes), além do formato OGG, da plataforma APPLE.

Vídeo No vídeo, estamos diante de necessidade maior de compressão. Se uma imagem estática já é grande, imagine precisar de 30 imagens estáticas por segundo. Agora, a localidade de referência ganha um novo viés, a l.r. no tempo. Ou seja, se imaginarmos um vídeo como um array 3-d (dimensões tempo, altura e largura), podemos comprimir em 2 dessas dimensões.

Aqui a história começa com os formatos analógicos (VHS, Betamax), segue pelo CD-ROM e VCD, e chega ao DVD. Depois vem os formatos AVC (padrão H.264), depois HEVC (H.265), VP9 (formato livre e aberto da Google) e AV1 (formato livre e aberto, melhor que o H.265). Diferentemente da imagem (e semelhantemente ao som), aqui existe muita preocupação com a latência (atraso no envio dos pacotes) que impacta diretamente a sincronia. O padrão H.264 utiliza timestamps e buffers para sincronizar áudio e vídeo. Em caso de perda de pacotes, mecanismos de recuperação de erros como o Concealment Motion Vector (CMV) podem ser utilizados para minimizar o impacto na sincronia. O CMV divide o vídeo em macroblocos. Quando há perda de um pacote contendo os vetores de movimento de um determinado macrobloco, o CMV busca os macroblocos vizinhos ($\uparrow, \downarrow, \leftarrow \rightarrow$) e ve se algum deles tem movimento similar ao esperado do macrobloco perdido. Daí ele "empresta" vetores e reconstitui o movimento ausente.

PDF O Portable Document Format (PDF), criado pela Adobe em 1993, se tornou um dos formatos de arquivo mais utilizados no mundo, presente em diversos setores e aplicações. Mais detalhes: vivx680a. Um documento PDF pode ser aberto e visualizado em qualquer dispositivo com um leitor de PDF instalado, independentemente do sistema operacional ou hardware utilizado. Isso garante que a formatação original do documento é sempre a mesma, independentemente da plataforma em que ele é aberto. Isso é crucial para documentos que exigem alta fidelidade visual, como apresentações, relatórios e publicações. O PDF oferece recursos de segurança robustos, como criptografia e assinaturas digitais, que protegem o conteúdo contra acessos não autorizados, modificações e falsificações. Isso torna o PDF ideal para documentos confidenciais e contratos legais. Até porque ele pode ser acompanhado de uma assinatura digital (*hash*). Suporta recursos de acessibilidade, como tags de estrutura e legendas para imagens, que facilitam o acesso ao conteúdo do documento por pessoas com deficiência visual ou outras necessidades especiais. Ao contrário de outros formatos de arquivo que podem ser corrompidos ou se tornar obsoletos com o tempo, o PDF é um formato durável e estável. Pense na economia de árvores: imagine se todos os PDFs existentes fossem para o papel. Ele é ideal para compartilhar documentos com outras pessoas, pois garante que a formatação original seja preservada e o conteúdo seja visualizado corretamente em diferentes dispositivos. O PDF é uma ótima opção para armazenar documentos importantes de forma segura e durável, pois protege o conteúdo contra modificações e garante que ele possa ser acessado no futuro. Revistas, livros, relatórios e outros materiais de publicação podem ser distribuídos no formato PDF, garantindo uma experiência de leitura consistente em diferentes dispositivos. **Outros:** Deixo de tratar aqui, por absoluta falta de espaço, outros BLOBs como exames médicos, arquivos de telescópios (vivx760), ...

Olhando dentro de um BLOB

Agora e hora de examinar com um "microscópio" as partes internas de um desses arquivos. O escolhido pela importância é o arquivo PDF. Pode parecer pouco, mas esta ideia simples, já salvou centenas de milhares (milhões ?) de árvores. Já são muitos os sistemas de informação que não geram mais saídas em papel, limitando-se a criar arquivos PDF.

Um arquivo PDF descreve como uma página ficará depois de impressa e ele é melhor do que imagem dessa página pelas seguintes razões

⋆ A imagem como tal sempre é muito maior do que

sua descrição. A razão para isso é a evidente redundância de qualquer imagem.

⋆ A imagem como tal é fixa e só pode ser rotacionada, redimensionada ou de alguma maneira modificada à custa de processamento e perda de qualidade.

⋆ Sobre uma imagem não há como manipular o conteúdo, por exemplo conduzindo uma busca textual (^F alguma coisa ou então ^C e ^V).

⋆ Alterar uma única palavra sobre uma imagem exige habilidades de pintura. Sobre a descrição de uma página é tarefa simples: basta trocar uma palavra pela outra.

⋆ Imagine uma imagem enviada para um monitor (72 DPP) ou a mesma imagem enviada para uma impressora (300 DPP). Não podem ser a mesma imagem.

A versão 1.0 do padrão PDF nasceu em 1993 quando a Adobe lançou 2 programas: o Distiller (para gerar PDF) e o Reader (para ler): ambos pagos. A coisa começou a mudar quando o departamento de rendas americano comprou uma licença que permitia aos contribuintes baixar o Reader. Rapidamente a Adobe deu-se conta de que seria melhor para ela distribuir gratuitamente o Reader. Nos dez anos seguintes o PDF acabou se tornando o padrão de fato de descrição de imagens.

Apresenta as seguintes vantagens:

⋆ acesso aleatório: qualquer página pode ser montada independente das demais.

⋆ linearização: todos os elementos necessários para montar uma página estão juntos.

⋆ atualização incremental: mudanças ficam registradas no fim do arquivo. Vantagem: opção de desfazer ilimitada e rápida atualização. Desvantagem: o arquivo não pára de crescer.

⋆ fontes incorporadas: o destino sempre será montado corretamente. A fonte é desbastada de tudo que não é necessário.

⋆ padronização ISO: em 2008 a ISO abraçou o padrão Adobe PDF 1.7, o que o transformou em padrão aberto.

⋆ compatibilidade para trás e para a frente: Qualquer leitor lê as versões antigas e deve ler as futuras, já que todos os leitores ignoram o que não conhecem.

⋆ integrado a todos os principais browsers do protocolo `HTTP`.

Um arquivo PDF

Cabeçalho Formado por 2 linhas: a primeira identifica o arquivo como um PDF e informa a versão PDF em que ele foi gerado. A segunda linha inclui caracteres que não podem ser impressos (para avisar eventuais leitores deste fato). Exemplo:

```
%PDF-1.0
%âã
```

Corpo Um conjunto de nodos de um grafo, contendo as páginas, conteúdo gráfico, textos, sempre na forma de objetos. Cada objeto começa com um número de objeto (iniciando em 1), um número de geração (sempre 0) e a palavra chave `obj`. O objeto termina pela palavra chave `endobj`. Exemplo:

```
1 0 obj     -- descrevendo o objeto 1
<<          -- começa um dicionário
 /Kids [2 0 R]     -- /Kids tem o valor 2 0 R
 /Count 1          -- só um
 /Type /Pages      -- do tipo : página
>>                 -- fim do dicionário
endobj             -- fim do objeto
```

Tabela de Referência Cruzada Lista o deslocamento em bytes de cada objeto no corpo do arquivo. Exemplo

```
0 6           -- seis entradas na tabela
0000000000  65535  f   -- entrada especial
0000000015  00000  n   -- obj1 no desloc 15
0000000074  00000  n   -- obj2 no desloc 74
...
```

Rodapé Informações e metadados para aumentar a performance de acesso. A primeira entrada é `trailer`. Depois o dicionário de trailer que deve apresentar ao menos a entrada `\Size` dizendo quantos itens há na tabela de referência cruzada e `\Root` dizendo qual objeto é o catálogo do documento. Depois vem uma linha que só contém `startxref` seguida por uma única linha indicando o deslocamento em bytes utilizado no inicio da tabela de referência cruzada do arquivo. Termina tudo a linha `%%EOF`. Exemplo:

```
trailer   -- palavra chave
<<   -- começa um dicionário
/Root 5 0 R   -- o catálogo é o objeto 5
/Size 6       -- tem o tamanho 6
>>   -- fim do dicionário
startxref
459  -- deslocamento da tabela de ref cruzada
%%EOF   -- fim de arquivo
```

Componentes

Uma descrição PDF suporta 5 componentes básicos: i. Números inteiros e reais (mas não exponenciais); ii. strings que são escritos entre parênteses; iii.Nomes, que sempre começam por uma barra normal; iv.Os valores booleanos `true` e `false` e v. O componente nulo (`null`). Aceita também 3 componentes compostos: i. Arrays, que são uma coleção ordenada de outros componentes, escritos entre colchetes; ii.Dicionários que são uma lista não ordenada de duplas: nome e componente. Começam por `<<` e terminam por `>>`. Finalmente iii. Fluxos, que armazenam dados binários, sempre acompanhados de um dicionário que descreve seus atributos, como comprimento e parâmetros de compactação, por exemplo.

☞ Para você fazer

As coisas em um arquivo PDF são separadas por um divisor de linhas. Pode ser o caracter $X'10'$, ou o $X'13'$ ou uma combinação de ambos. Ao examinar um PDF em hexadecimal, acostume-se a separar as linhas marcando este caractere. Isto posto, os componentes principais do arquivo PDF são:

- Header: identificação PDF e uma coleção de letras acentuadas
- Corpo: objetos numerados a partir de 1 sequencialmente, e identificados pela palavra `obj` no início e `endobj` no final. Esses objetos podem ser textos, desenhos, fontes, imagens, metadados, links, e um monte de coisas mais.
- Tabela de referência: uma lista de onde (no arquivo) está cada objeto. Esta tabela permite o acesso direto a cada um e a todos os objetos.
- Rodapé: Diversas informações que permitem o acesso rápido ao arquivo.

Nos interessa, neste exercício a informação `startxref` que é uma das últimas do arquivo. Você deve abrir o arquivo

```
arq08.pdf
```

publicado no lugar usual, com um editor de hexadecimal e deve procurar o startxref no final do arquivo. Na linha seguinte, existe um endereço que o PDF coloca em decimal. Você precisa transformar este número em hexadecimal e localizar no arquivo (mais acima) este endereço.

Nele, deve aparecer a palavra `xref` e na linha seguinte um zero e um número. Na linha seguinte, uma entrada (que todo arquivo tem) com muitos zeros, o número 65535 e uma letra. Na linha seguinte, **um número** seguido de zeros e uma letra. Trata-se do endereço do primeiro objeto do arquivo.

Responda aqui:

endereço da xref em hexadecimal	endereço do primeiro objeto em decimal

CEP - UP - UTFPR - PUCPr - UFPR –
17/06/2024 - 17:11:23.4
Prof Dr P Kantek (pkantek@gmail.com)
Objetos binários grandes vivxq10a, V: 1.01 76656
VITOR EDUARDO DE GOES FONTES
24CC2301 - 17 entregar ate 4/julho _____ /
/ _____

Objetos binários grandes - BLOBs

O objetivo desta atividade e conhecer e manusear os chamados objetos binários grandes: imagens, vídeos, documentos, músicas, etc. Tais objetos quando estão na memória do computador são conhecidos como BLOBs e quando vão para uma memória não volátil (pendrives, discos, SSDs, nuvem, ...) passam a ser arquivos.

Resultados da digitalização crescente da nossa sociedade, precisam ser conhecidos para deles se extraírem todos os benefícios que eles apresentam.

Eis alguns raciocínios: a documentação em papel de um avião boeing 747 ocuparia 5 aviões 747 para ser transportada. Alguém ainda compra enciclopédias em papel ? Lembram de quando as fotografias eram analógicas ? Um filme Ektachrome (36 fotos) custava bem caro. Vou chutar uns 150 reais, o que daria quase 5 reais/foto, só o filme, sem contar a revelação e as ampliações. Quando o I.R. era em papel, a receita precisava montar uma operação de guerra, envolvendo todos os bancos, apenas para receber as declarações, e isso que era mais ou menos 1/3 do número de contribuintes atual. E as músicas: durante muitos anos, no dia do meu pagamento eu ia a uma loja de discos (?) e comprava 1 único LP, ($\pm$ 30 minutos de música) era o que o meu dinheiro alcançava. Aliás, falando de música, vale comentar a ascensão e queda do formato CD. Ele passou de maravilha tecnológica em meados dos 80 a algo obsoleto em meados dos 2010. Do céu ao inferno em 30 anos. E, se perder em uma cidade estranha, tendo apenas o guia 4 Rodas para ajudar ? Uma sensação indescritível.

Vamos estudar mais de perto alguns BLOBs, lembrando que qualquer um pode criar e chamar de seu, um BLOB. Pela sua definição (grande bloco de dados binários com uma lógica interna de organização e acesso) vê-se que isto é perfeitamente possível. Assim, vamos nos limitar a BLOBs padrão como formatos e conceitos conhecidos e publicados.

Imagem É difícil contar esta história. Ela começa no daguerrótipo de meados do século 19, através da fixação de imagens em placas de vidro. Segue pelas imagens de TV que eram capturadas e transmitidas eletronicamente. Seu registro foi feito em gravações magnéticas analógicas. Uma nova necessidade surgiu na obtenção de fotos na Lua e além, quando não havia como retornar filmes. A propósito, um pouco antes disto, a vigilância eletrônica na Guerra Fria era muito custosa: satélites fotografavam o inimigo e precisavam ejetar os filmes (que acabavam rápido) ao sobrevoar território amigo. A primeira foto da lua, transmitida eletronicamente foi feita pela espaçonave Ranger em 1964, 17 minutos antes dela se espatifar na superfície lunar. Quando a Internet começou (por volta de 1990) a necessidade de manusear imagens sofreu uma mudança: antes o gerador e o visualizador eram de mesma origem, pelo que não havia necessidade de padronização. Com a Internet essa história mudou. Uma das primeiras respostas a esta demanda foi o formato BMP (Bitmap Image File). Embora originalmente proposto pela Microsoft e IBM, teve seu padrão publicado e virou de fato, um formato livre. Sua origem é a bio-física dos olhos humanos (3 canais separados para Red, Green e Blue), além de um truque bem legal chamado mapeamento pelo qual se negocia tamanho do arquivo VERSUS menor quantidade de cores. A principal inovação do BMP foi criar, publicar e obedecer a um padrão. Parece pouco, mas não é. A chave da popularização e do sucesso sempre é um PADRãO. Anote isso na sua cabeça e nunca mais esqueça. O padrão pode ser ruim ou bom (neste caso não é muito), mas é um padrão. A imagem é uma matriz numérica (outra inovação entusiasmadora), logo sujeita a sofrer a ação maravilhosa de qualquer matemática.

Padrão BMP O arquivo começa com um bloco de controle de 54 bytes, que contém o identificador 'BM', o tamanho do objeto (arquivo), onde a imagem começa, L=largura e A=altura da imagem, profundidade). Depois, dependendo da profundidade, pode haver uma tabela de cores. Finalmente a imagem na forma de $L \times A$ se for uma imagem mapeada ou 3 tabelas de $L \times A$, uma para cada cor se for uma imagem true-color. Veja na imagem a seguir estes conceitos

```
Exemplo:
        0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
       -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
0000   42 4D 62 05 00 00 00 00 00 00 36 04 00 00 28 00 BM.....6...(
0010   00 00 12 00 00 00 0F 00 00 00 01 00 08 00 00 00 ................
0020   00 00 60 01 00 00 00 00 00 00 00 00 00 00 00 01 ................
0030   00 00 00 00 00 00 00 00 FF FF 00 80 80 80 00 FF FF ............
0040   FF 00 FF 00 FF 00 00 00 FF 00 00 00 00 00 00 FF ................
0050   00 00 FF FF 00 00 FF 00 00 00 09 09 09 00 0A 0A ................
0060   0A 00 0B 0B 0B 00 0C 0C 0C 00 0D 0D 0D 00 0E 0E ................
0070   0E 00 0F 0F 0F 00 10 10 10 00 11 11 11 00 12 12 ................
...
0410   F6 00 F7 F7 F7 00 F8 F8 F8 00 F9 F9 F9 00 FA FA ................
0420   FA 00 FB FB FB 00 FC FC FC 00 FD FD FD 00 FE FE ................
0430   FE 00 FF FF FF 00 01 01 03 01 01 01 08 01 01 01 ................
0440   01 01 02 01 01 06 01 01 00 00 04 04 03 04 04 04 ................
0450   04 04 04 04 04 04 04 04 04 04 04 04 00 00 01 01 ................
```

Os problemas do BMP incluem: nenhuma compressão, o que graças ao fenômeno da *localidade de referência espacial* é uma tragédia. Também não prevê nenhum tipo de animação, o que faz falta na Internet. Não sabia (hoje sabe) lidar com transparência.

Um concorrente importante do BMP (lá nos primórdios) foi o formato GIF que admite compressão (usando o belo algoritmo LZW) e também admite uma forma rudimentar de animação. O problema do GIF é que ele foi e é um formato proprietário. Isto deu origem ao formato PNG que é praticamente o mesmo, mas livre.

A compressão poderia ser muito melhor, se se admitisse algum nivel de degradação na reconstituição da imagem após sua transformação em BLOB. Isto foi conseguido no padrão JPEG, que ao admitir degradações (sob controle), gera compressões muito maiores do que GIF ou similares. Outra vantagem é a possibilidade de criação em etapas das imagens (coisa que o BMP nunca ofereceu). O padrão foi formalizado pela ISO em 1991, mas na época disparou uma briga por patentes e direitos. Nos EUA, as últimas patentes expiraram em 2004. A compressão aqui é pelo uso da Transformada Discreta do Coseno. A sua operação está além do nosso interesse (querendo, olhe VIVX690c ou 690p), mas em resumo: a imagem é quebrada em blocos de 8×8; esta matriz é multiplicada pela matriz DCT, o que passa do domínio do pixel para o domínio da frequência (Transformada de Fourier); os coeficientes são quantizados; os coeficientes são comprimidos. Depois, se faz o inverso e *voilà*.

Som O MP3, sigla para MPEG-1 Audio Layer III, é um formato de arquivo de áudio digital que revolucionou a forma como consumimos e compartilhamos música.

Desenvolvido no início da década de 1990 pelo Moving Picture Experts Group (MPEG), o MP3 se tornou o formato dominante para música digital, utilizado em players de música portáteis, computadores, smartphones e serviços de streaming online. O MP3 utiliza uma técnica de compressão com perda chamada percepção auditiva psicoacústica para reduzir o tamanho dos arquivos de áudio. Essa técnica remove informações da faixa de áudio que o ouvido humano provavelmente não consegue perceber, resultando em arquivos menores sem comprometer significativamente a qualidade da música para a maioria dos ouvintes. Apesar da compressão, o MP3 oferece uma qualidade de som bastante boa, especialmente em taxas de bits mais altas (como 128 kbps ou superior). Tamanho de Arquivo Reduzido: Os arquivos MP3 são significativamente menores do que os arquivos de áudio não compactados, como CDs, permitindo fácil armazenamento e compartilhamento de música digital.

O MP3 teve um impacto profundo na indústria da música e na forma como consumimos música gerando um declínio das vendas de CDs: Os consumidores passaram a baixar e compartilhar músicas digitalmente. O MP3 impulsionou o crescimento da música digital, com o surgimento de serviços de streaming online como Spotify, Apple Music e Deezer. O MP3 abriu caminho para novos modelos de negócios na indústria da música, como a venda de músicas online e assinaturas de serviços de streaming.

Apesar de seu sucesso, o MP3 também enfrentou alguns desafios: Em taxas de bits mais baixas (como 64 kbps), a qualidade de som do MP3 pode ser perceptívelmente inferior à do áudio não compactado. Propriedades Intelectuais: Questões de direitos autorais e pirataria surgiram com a proliferação de arquivos MP3 compartilhados online. Formatos Alternativos: Novos formatos de áudio digital, como FLAC e AAC, oferecem maior qualidade de som ou taxas de compressão mais eficientes, mas ainda não alcançaram o mesmo nível de popularidade do MP3.

Outros padrões: MID (notação musical sintética), WAV (nenhuma perda de qualidade, logo arquivos grandes), além do formato OGG, da plataforma APPLE.

Vídeo No vídeo, estamos diante de necessidade maior de compressão. Se uma imagem estática já é grande, imagine precisar de 30 imagens estáticas por segundo. Agora, a localidade de referência ganha um novo viés, a l.r. no tempo. Ou seja, se imaginarmos um vídeo como um array 3-d (dimensões tempo, altura e largura), podemos comprimir em 2 dessas dimensões.

Aqui a história começa com os formatos analógicos (VHS, Betamax), segue pelo CD-ROM e VCD, e chega ao DVD. Depois vem os formatos AVC (padrão H.264), depois HEVC (H.265), VP9 (formato livre e aberto da Google) e AV1 (formato livre e aberto, melhor que o H.265). Diferentemente da imagem (e semelhantemente ao som), aqui existe muita preocupação com a latência (atraso no envio dos pacotes) que impacta diretamente a sincronia. O padrão H.264 utiliza timestamps e buffers para sincronizar áudio e vídeo. Em caso de perda de pacotes, mecanismos de recuperação de erros como o Concealment Motion Vector (CMV) podem ser utilizados para minimizar o impacto na sincronia. O CMV divide o vídeo em macroblocos. Quando há perda de um pacote contendo os vetores de movimento de um determinado macrobloco, o CMV busca os macroblocos vizinhos ($\uparrow, \downarrow, \leftarrow \rightarrow$) e ve se algum deles tem movimento similar ao esperado do macrobloco perdido. Daí ele "empresta" vetores e reconstitui o movimento ausente.

PDF O Portable Document Format (PDF), criado pela Adobe em 1993, se tornou um dos formatos de arquivo mais utilizados no mundo, presente em diversos setores e aplicações. Mais detalhes: vivx680a. Um documento PDF pode ser aberto e visualizado em qualquer dispositivo com um leitor de PDF instalado, independentemente do sistema operacional ou hardware utilizado. Isso garante que a formatação original do documento é sempre a mesma, independentemente da plataforma em que ele é aberto. Isso é crucial para documentos que exigem alta fidelidade visual, como apresentações, relatórios e publicações. O PDF oferece recursos de segurança robustos, como criptografia e assinaturas digitais, que protegem o conteúdo contra acessos não autorizados, modificações e falsificações. Isso torna o PDF ideal para documentos confidenciais e contratos legais. Até porque ele pode ser acompanhado de uma assinatura digital (*hash*). Suporta recursos de acessibilidade, como tags de estrutura e legendas para imagens, que facilitam o acesso ao conteúdo do documento por pessoas com deficiência visual ou outras necessidades especiais. Ao contrário de outros formatos de arquivo que podem ser corrompidos ou se tornar obsoletos com o tempo, o PDF é um formato durável e estável. Pense na economia de árvores: imagine se todos os PDFs existentes fossem para o papel. Ele é ideal para compartilhar documentos com outras pessoas, pois garante que a formatação original seja preservada e o conteúdo seja visualizado corretamente em diferentes dispositivos. O PDF é uma ótima opção para armazenar documentos importantes de forma segura e durável, pois protege o conteúdo contra modificações e garante que ele possa ser acessado no futuro. Revistas, livros, relatórios e outros materiais de publicação podem ser distribuídos no formato PDF, garantindo uma experiência de leitura consistente em diferentes dispositivos. **Outros:** Deixo de tratar aqui, por absoluta falta de espaço, outros BLOBs como exames médicos, arquivos de telescópios (vivx760), ...

Olhando dentro de um BLOB

Agora e hora de examinar com um "microscópio" as partes internas de um desses arquivos. O escolhido pela importância é o arquivo PDF. Pode parecer pouco, mas esta ideia simples, já salvou centenas de milhares (milhões ?) de árvores. Já são muitos os sistemas de informação que não geram mais saídas em papel, limitando-se a criar arquivos PDF.

Um arquivo PDF descreve como uma página ficará depois de impressa e ele é melhor do que imagem dessa página pelas seguintes razões

⋆ A imagem como tal sempre é muito maior do que

sua descrição. A razão para isso é a evidente redundância de qualquer imagem.

⋆ A imagem como tal é fixa e só pode ser rotacionada, redimensionada ou de alguma maneira modificada à custa de processamento e perda de qualidade.

⋆ Sobre uma imagem não há como manipular o conteúdo, por exemplo conduzindo uma busca textual (^F alguma coisa ou então ^C e ^V).

⋆ Alterar uma única palavra sobre uma imagem exige habilidades de pintura. Sobre a descrição de uma página é tarefa simples: basta trocar uma palavra pela outra.

⋆ Imagine uma imagem enviada para um monitor (72 DPP) ou a mesma imagem enviada para uma impressora (300 DPP). Não podem ser a mesma imagem.

A versão 1.0 do padrão PDF nasceu em 1993 quando a Adobe lançou 2 programas: o Distiller (para gerar PDF) e o Reader (para ler): ambos pagos. A coisa começou a mudar quando o departamento de rendas americano comprou uma licença que permitia aos contribuintes baixar o Reader. Rapidamente a Adobe deu-se conta de que seria melhor para ela distribuir gratuitamente o Reader. Nos dez anos seguintes o PDF acabou se tornando o padrão de fato de descrição de imagens.

Apresenta as seguintes vantagens:

⋆ acesso aleatório: qualquer página pode ser montada independente das demais.

⋆ linearização: todos os elementos necessários para montar uma página estão juntos.

⋆ atualização incremental: mudanças ficam registradas no fim do arquivo. Vantagem: opção de desfazer ilimitada e rápida atualização. Desvantagem: o arquivo não pára de crescer.

⋆ fontes incorporadas: o destino sempre será montado corretamente. A fonte é desbastada de tudo que não é necessário.

⋆ padronização ISO: em 2008 a ISO abraçou o padrão Adobe PDF 1.7, o que o transformou em padrão aberto.

⋆ compatibilidade para trás e para a frente: Qualquer leitor lê as versões antigas e deve ler as futuras, já que todos os leitores ignoram o que não conhecem.

⋆ integrado a todos os principais browsers do protocolo `HTTP`.

Um arquivo PDF

Cabeçalho Formado por 2 linhas: a primeira identifica o arquivo como um PDF e informa a versão PDF em que ele foi gerado. A segunda linha inclui caracteres que não podem ser impressos (para avisar eventuais leitores deste fato). Exemplo:

```
%PDF-1.0
%âã
```

Corpo Um conjunto de nodos de um grafo, contendo as páginas, conteúdo gráfico, textos, sempre na forma de objetos. Cada objeto começa com um número de objeto (iniciando em 1), um número de geração (sempre 0) e a palavra chave `obj`. O objeto termina pela palavra chave `endobj`. Exemplo:

```
1 0 obj     -- descrevendo o objeto 1
<<          -- começa um dicionário
 /Kids [2 0 R]     -- /Kids tem o valor 2 0 R
 /Count 1          -- só um
 /Type /Pages      -- do tipo : página
>>                 -- fim do dicionário
endobj             -- fim do objeto
```

Tabela de Referência Cruzada Lista o deslocamento em bytes de cada objeto no corpo do arquivo. Exemplo

```
0 6          -- seis entradas na tabela
0000000000  65535  f   -- entrada especial
0000000015  00000  n   -- obj1 no desloc 15
0000000074  00000  n   -- obj2 no desloc 74
...
```

Rodapé Informações e metadados para aumentar a performance de acesso. A primeira entrada é `trailer`. Depois o dicionário de trailer que deve apresentar ao menos a entrada `\Size` dizendo quantos itens há na tabela de referência cruzada e `\Root` dizendo qual objeto é o catálogo do documento. Depois vem uma linha que só contém `startxref` seguida por uma única linha indicando o deslocamento em bytes utilizado no inicio da tabela de referência cruzada do arquivo. Termina tudo a linha `%%EOF`. Exemplo:

```
trailer   -- palavra chave
<<   -- começa um dicionário
/Root 5 0 R   -- o catálogo é o objeto 5
/Size 6       -- tem o tamanho 6
>>   -- fim do dicionário
startxref
459  -- deslocamento da tabela de ref cruzada
%%EOF   -- fim de arquivo
```

Componentes

Uma descrição PDF suporta 5 componentes básicos: i. Números inteiros e reais (mas não exponenciais); ii. strings que são escritos entre parênteses; iii.Nomes, que sempre começam por uma barra normal; iv.Os valores booleanos `true` e `false` e v. O componente nulo (`null`). Aceita também 3 componentes compostos: i. Arrays, que são uma coleção ordenada de outros componentes, escritos entre colchetes; ii.Dicionários que são uma lista não ordenada de duplas: nome e componente. Começam por `<<` e terminam por `>>`. Finalmente iii. Fluxos, que armazenam dados binários, sempre acompanhados de um dicionário que descreve seus atributos, como comprimento e parâmetros de compactação, por exemplo.

☞ Para você fazer

As coisas em um arquivo PDF são separadas por um divisor de linhas. Pode ser o caracter $X'10'$, ou o $X'13'$ ou uma combinação de ambos. Ao examinar um PDF em hexadecimal, acostume-se a separar as linhas marcando este caractere. Isto posto, os componentes principais do arquivo PDF são:

- Header: identificação PDF e uma coleção de letras acentuadas
- Corpo: objetos numerados a partir de 1 sequencialmente, e identificados pela palavra `obj` no início e `endobj` no final. Esses objetos podem ser textos, desenhos, fontes, imagens, metadados, links, e um monte de coisas mais.
- Tabela de referência: uma lista de onde (no arquivo) está cada objeto. Esta tabela permite o acesso direto a cada um e a todos os objetos.
- Rodapé: Diversas informações que permitem o acesso rápido ao arquivo.

Nos interessa, neste exercício a informação `startxref` que é uma das últimas do arquivo. Você deve abrir o arquivo

```
arq01.pdf
```

publicado no lugar usual, com um editor de hexadecimal e deve procurar o startxref no final do arquivo. Na linha seguinte, existe um endereço que o PDF coloca em decimal. Você precisa transformar este número em hexadecimal e localizar no arquivo (mais acima) este endereço.

Nele, deve aparecer a palavra `xref` e na linha seguinte um zero e um número. Na linha seguinte, uma entrada (que todo arquivo tem) com muitos zeros, o número 65535 e uma letra. Na linha seguinte, **um número** seguido de zeros e uma letra. Trata-se do endereço do primeiro objeto do arquivo.

Responda aqui:

endereço da xref em hexadecimal	endereço do primeiro objeto em decimal