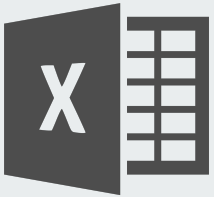
Live de Python #25

Raspando arquivos não HTML



Roteiro

- Como funcionam os documentos com extensão livre? (ODF)
- Como funcionam os documentos com extensão proprietárias? (MS Office)
- Como funcionam os PDFs?
- Lib que vamos usar:
  - ezodf (ODF)
  - ??? (MS files)
  - pdfminner (PDF)

Senta que lá vem a história dos ODs

OD/ODF (Open Document File) são documentos de formato aberto para arquivos de escritório (planilhas, apresentações, texto, que foram padrões desenvolvidos pela **OASIS** (Organization for the advancement of structured information standards) e é um padrão aberto.

https://pt.wikipedia.org/wiki/OpenDocument
https://pt.wikipedia.org/wiki/OASIS_(organiza%C3%A7%C3%A3o)

Especificações dos ODF

- .odt e .fodt para documentos de texto (text)
- .ods e .fods para folhas de cálculo (spreadsheets)
- .odp e .fodp para apresentações (presentations)
- .odb para base de dados (database)
- .odg e .fodg para desenhos vetoriais (graphic)
- .odf para equações (formula)
- .ott para modelos de documentos de texto (template text)
- .ots para modelos de folhas de cálculo (template spreadsheets)
- .otp para modelos de apresentações (template presentations)
- .otg para modelos de desenhos vetoriais (template graphic)

Quem usa ODF como padrão de documentos?

- África do Sul
- Angola
- Alemanha
- Bélgica
- Brasil
- Croácia
- Dinamarca
- Equador
- França
- Holanda
- Japão
- Malásia
- Noruega
- Polônia
- Portugal
- Reino Unido
- Rússia
- Suécia
- Uruguai
- Venezuela

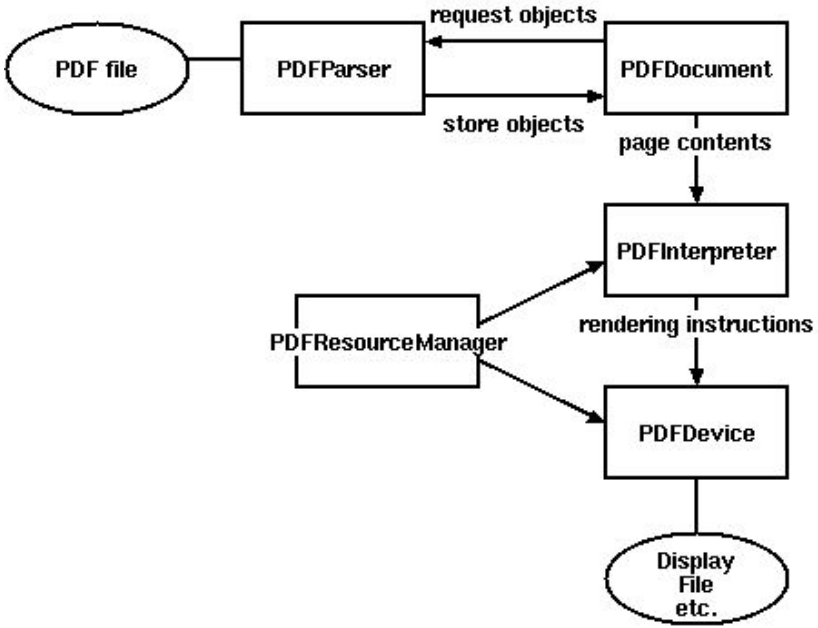
Quem usa ODF no Brasil? (órgãos públicos)

- Banco do Brasil
- Caixa Econômica Federal
- (Celepar)
- CONAB
- Ministério Público Federal
- Correios
- Dataprev
- Exército Brasileiro
- Aeronáutica
- IBGE
- INPE
- INPI
- INPI
- Itaipu Binacional
- ITI
- Jardim Botânico de Brasília
- Marinha do Brasil
- Ministério do Planejamento
- Ministério das Relações Exteriores
- Ministério da Justiça
- Petrobras
- Prefeitura Municipal de Novo Hamburgo
- Prefeitura Municipal de Silva Jardim
- Tribunal Regional do Trabalho da 4ª Região
- PTI - Fundação Parque Tecnológico de Itaipu
- PRODERJ
- SERPRO
- SLTI
- UNESP
- UFV

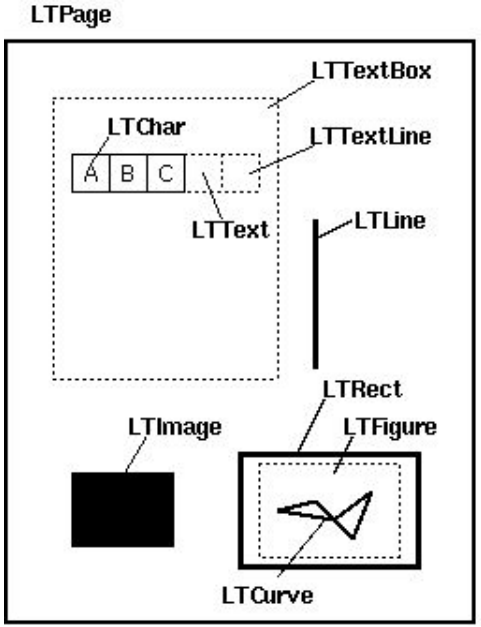
Como funcionam as extensões proprietárias?

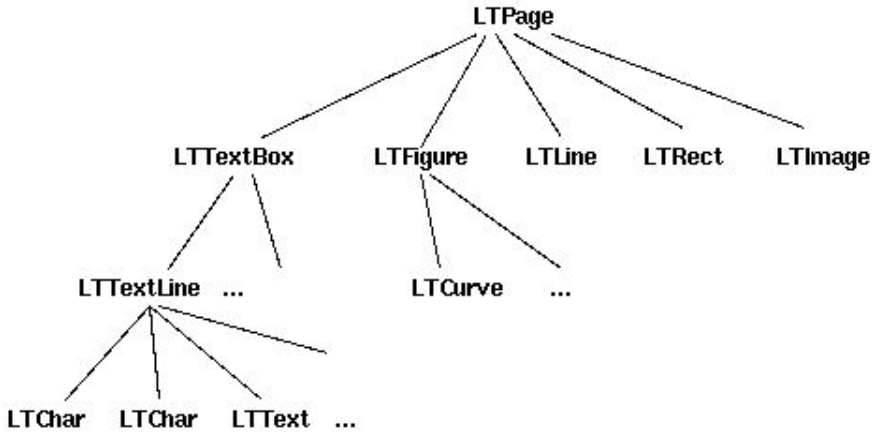
Não sei.....

PDFs

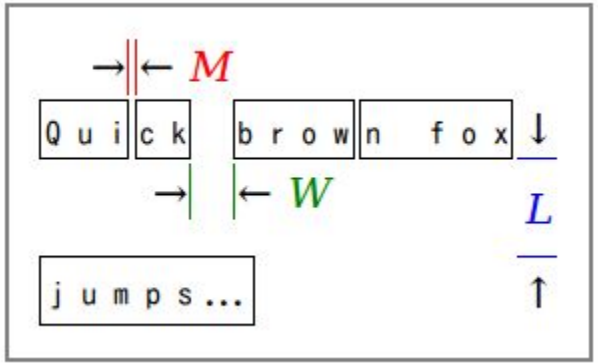

PDF file
PDFParser
request objects
store objects
PDFDocument
page contents
PDFInterpreter
PDFResourceManager
rendering instructions
PDFDevice
Display
File
etc.

PDFs


LTPage
LTTextBox
LTChar
LTTextLine
A
B
C
LTText
LTLine
LTRect
LTImage
LTFigure
LTCurve


LTPage
LTTextBox
LTFigure
LTLine
LTRect
LTImage
LTTextLine ...
LTCurve ...
LTChar
LTChar
LTText ...

PDFs



http://pdfminer-docs.readthedocs.io/programming.html

ezodf

http://pythonhosted.org/ezodf/