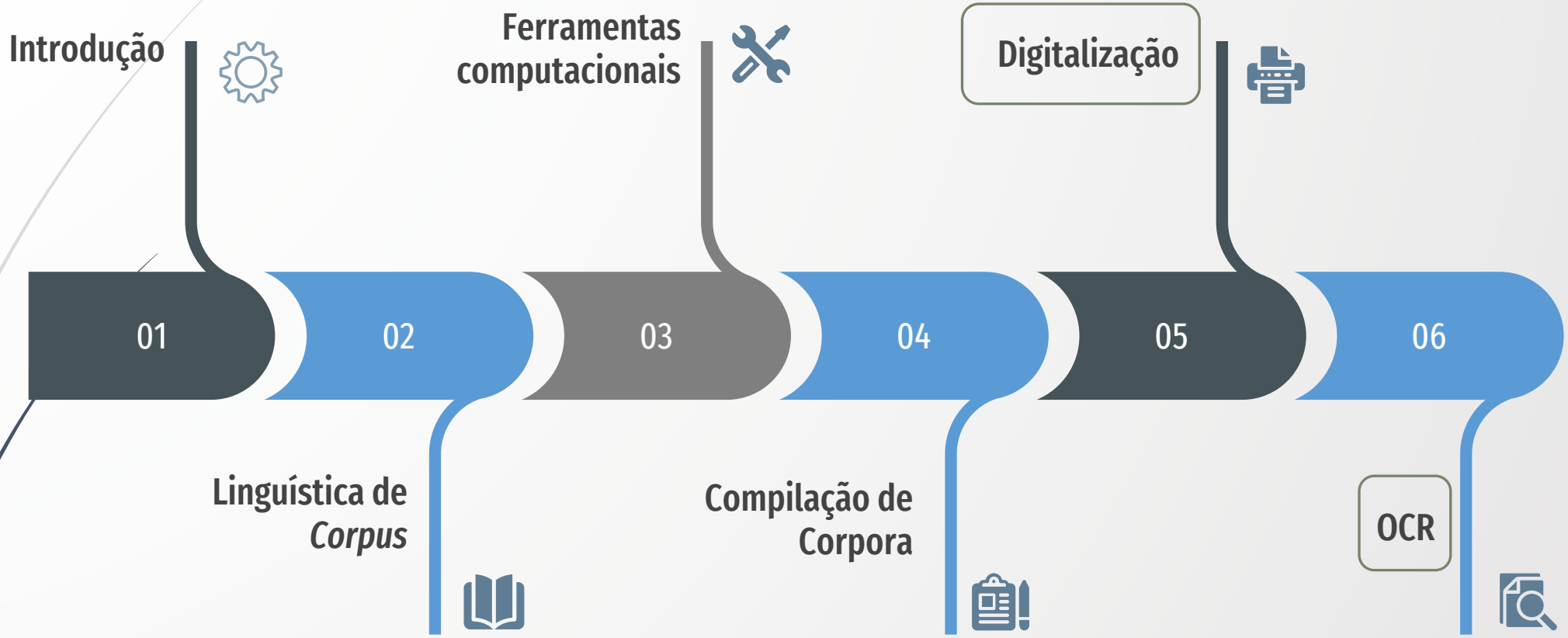
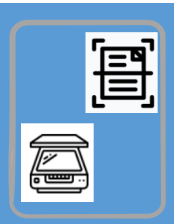
Humanities Going Digital (HUGOD)

[2020-1-CZ01-KA226-HE-094363]

Compilação de Corpora - OCR	Sessão 04	12/10/2022	Chiara Barbero Sílvia Barbosa

AULAS


Introdução
01
Linguística de *Corpus*
02
Ferramentas computacionais
03
Compilação de Corpora
04
Digitalização
05
OCR
06



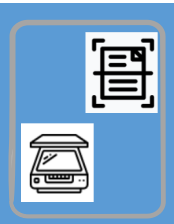
OCR (*Optical Character Recognition* | Reconhecimento Óptico de Caracteres)

O que é o OCR?

- O OCR permite a conversão de documentos em diversos formatos (GIF, JPG, PNG, TIFF e PDF) em dados em formato de texto que podem ser pesquisados e editados.

Como funciona?

- A ferramenta analisa o documento, compara os caracteres com fontes armazenadas e/ou reconhece características típicas de determinado caractere.



OCR (*Optical Character Recognition* | Reconhecimento Óptico de Caracteres)

Como funciona o reconhecimento do OCR?

- compara cada caractere identificado previamente com uma base de símbolos para definir padrões e encontrar semelhanças.
- depois da identificação e definição dos caracteres, o OCR compara as informações extraídas com uma base de palavras da língua em questão, de forma a confirmar ou não os dados extraídos.

Quais as vantagens?

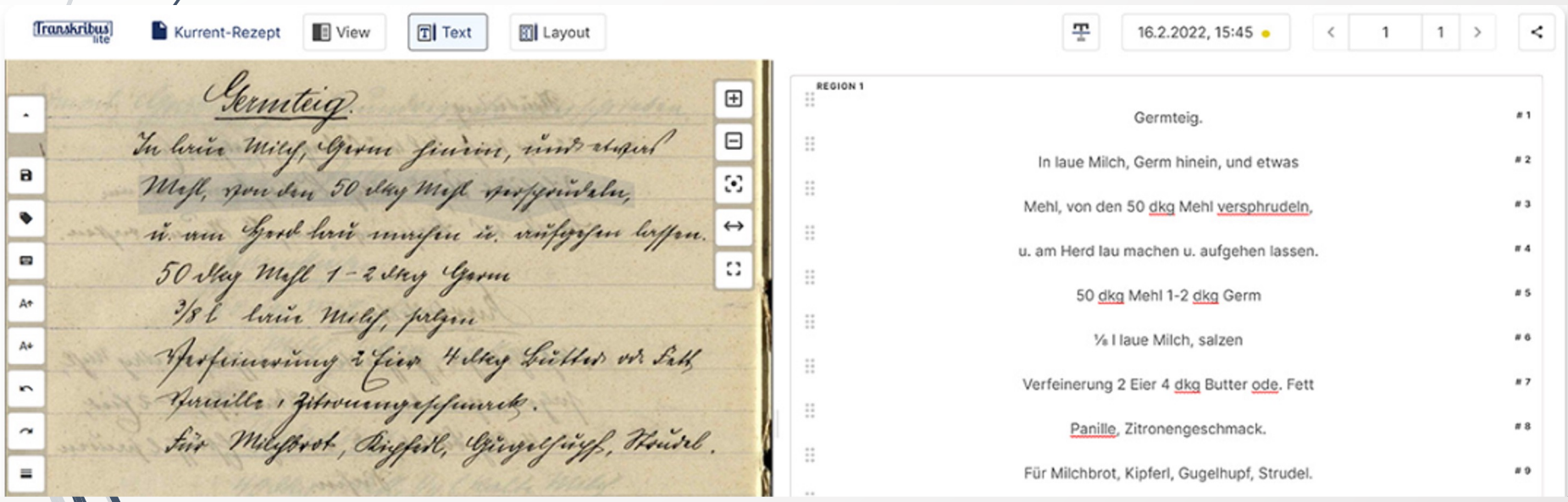
- permitir o acesso completo ao textos para pesquisa, edição e armazenamento
- reduzir ao mínimo a intervenção manual
- possibilitar a conversão para diferentes tipos de formatos

HWR / HTR

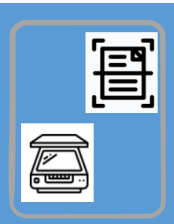
(Handwritten Recognition / Handwritten Text / Recognition | Reconhecimento Óptico de Caracteres)

O que é o HWR/HTR?

- Ferramenta para reconhecimento de textos manuscritos
- Utilizada em projetos que lidem com documentação histórica (cartas, diários, …)
- https://readcoop.eu/transkribus/



Transkribus Explained



OCR e *Corpora*

Quando recorrer ao uso de OCR para a compilação de *corpora*?

- Quando for necessário recorrer à digitalização, para a transformação das "imagens" até se tornarem "textos"

- Na consulta e uso de materiais presentes em repositórios em formato eletrónico mas não editável (muito frequente!!)

- Porque a maioria das ferramentas para exploração de *corpora* exige textos em formato editável, maioritariamente em formato de texto simples (.txt)

Captura de Imagem

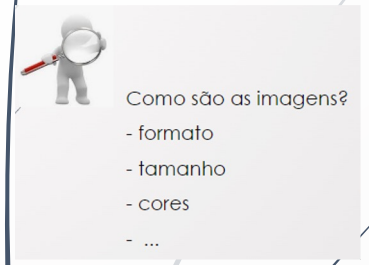
Como são as imagens?

- formato
- tamanho
- cores
- ...

Que ferramentas?

- digitalizador/*scanner*
- máquina fotográfica
- telemóvel/*smartphone*

Captura de Imagem



Formatos

- imagem com texto a cores (TIFF/JPEG2000, 300 ppi resolução, 24bit, sem compressão)
- imagem com texto p/b (TIFF/JPEG2000, 300 ppi resolução, 8bit, sem compressão)
- JPEG/PNG também são possíveis

Otimização da imagem obtida (alterações para melhor OCR)

- corte das áreas não necessárias
- orientação vertical/horizontal (para melhorar a leitura)
- formatação (das cores, da tonalidade, das curvaturas)
- limpeza de marcas (borrões, marcas, outros problemas)

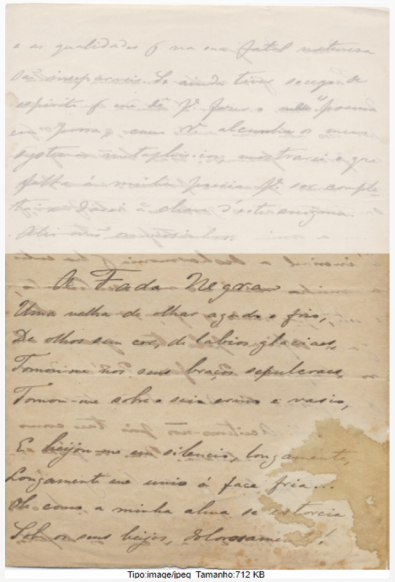
Gravação da nova imagem

- descrição dos procedimentos usados para replicar nas seguintes no formato desejado

Captura de Imagem: exemplos

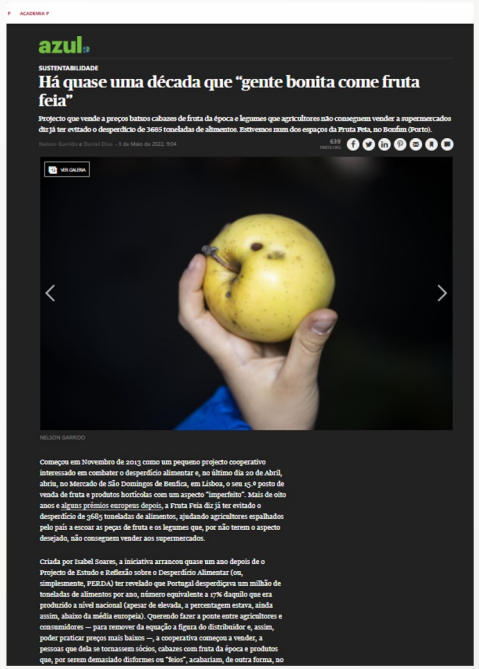
Existe um conjunto diversificado de material em diferentes formatos que pode ser alvo de digitalização

A



Tipo:image/jpeg Tamanho:712 KB

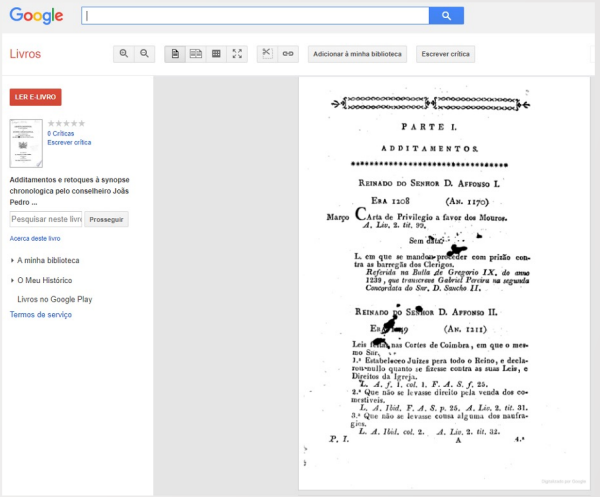
B



C

D

E



F

Que ferramenta usar?

Como escolher?

- De acordo com o objeto - documento manuscrito (HTR) ou ficheiro eletrónico (OCR)
- De acordo com o formato de ficheiro eletrónico (imagem ou PDF)
- De acordo com a língua de trabalho
- De acordo com a formatação do texto (texto corrido, diferentes colunas..)
- De acordo com a variedade linguística em uso (contemporânea ou grafias arcaicas) e o alfabeto (cirílico, hebraico, etc.) ou sistema de caracteres (japonês, coreano, etc.)
- De acordo com as licenças à nossa disposição (softwares pagos)

Testar sempre uma amostra e avaliar os resultados para escolher a ferramenta que melhor se adapta à nossa investigação

Tratamento da imagem

Programas/sotware de HTR

- Transkribus: (https://readcoop.eu/transkribus/)
- Amazon Textract (https://aws.amazon.com/pt/textract/)

Programas / software de OCR

Pagos

- Adobe Acrobat Pro DC
- PDF Reader
- OmniPage Ultimate
- ABBYY FineReader PDF
- SimpleOCR
- Tesseract

Gratuitos

- Online OCR (https://www.onlineocr.net/pt/)
- Tools PDF24 (https://tools.pdf24.org/pt/ocr-pdf)
- OCR Space (https://ocr.space/)
- OCR 2 Edit (https://www.ocr2edit.com/pt/converter-para-txt)
- New OCR (https://www.newocr.com/)

Avaliação dos resultados/problemas mais frequentes

Problemas de ordem gráfica:

- falta de sinais diacríticos (ou sinais diacríticos manipulados, acentuação errada)
- falta de caracteres
- caracteres a mais que não existem no original
- falta de espaços brancos entre duas palavras
- espaços brancos a mais no meio das palavras
- hifenização alterada

Problemas de formatação:

- desformatação geral do texto
- manipulação de metadados (desaparecem ou acrescentar marcas do *software*)
- reorganização do texto (no caso de textos com mais de uma coluna) que não respeita a ordem original

Trabalho em aula – apresentar no dia 17 os resultados

- (1) Selecionar a imagem (1 imagem por grupo) / (2) selecionar um conjunto de imagens
- (1) Testar a imagem com *softwares* diferentes / (2) testar todas as imagens do conjunto selecionado num dos software propostos
- Observar os resultados (pontos forte e os pontos fracos)
- Escolher o *software* que melhor se adapta
- Anotar o procedimento metodológico

IMAGEM

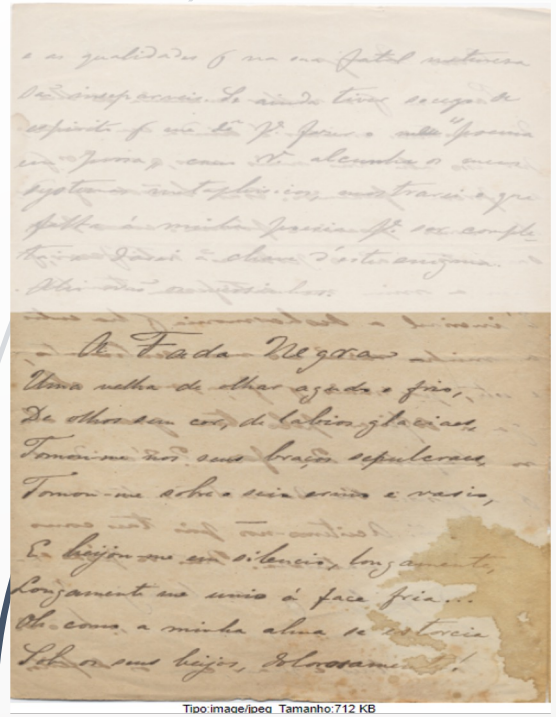
? Formato, ? tamanho, ? Cores, ? Caracteres, …

TEXTO EDITÁVEL
(resultado é um texto num ficheiro com extensão .TXT)

"O choco de Setúbal não é o único a estar frito."

EXEMPLOS DE IMAGENS

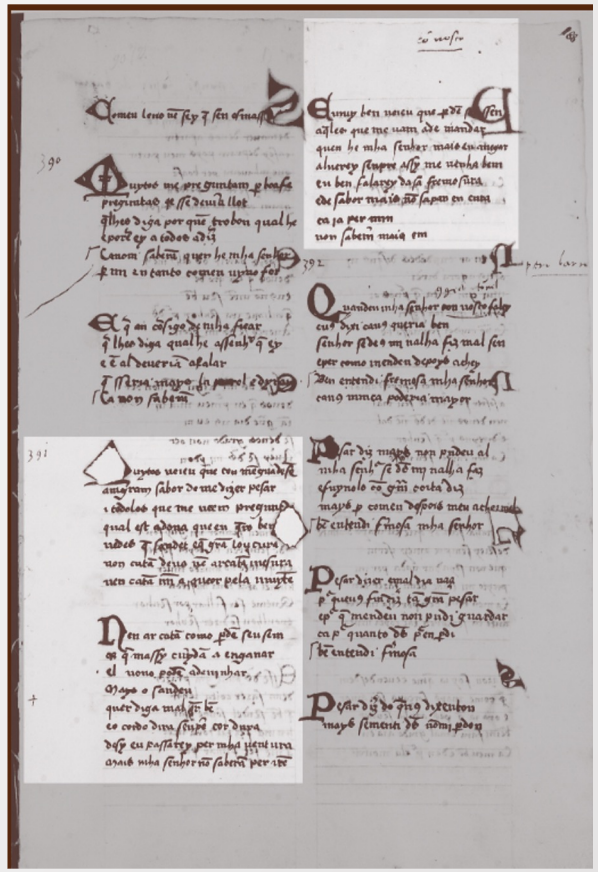
- Existe um conjunto diversificado de material em diferentes formatos que pode ser alvo de digitalização



G1

G2

G3



EXEMPLOS DE IMAGENS

- Existe um conjunto diversificado de material em diferentes formatos que pode ser alvo de digitalização

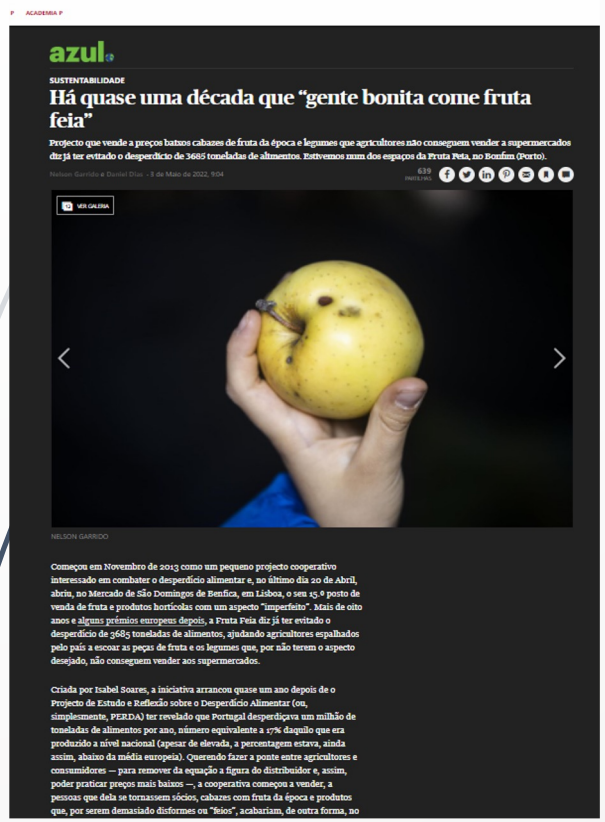
G4

G5

G6

EXEMPLOS DE IMAGENS

- Existe um conjunto diversificado de material em diferentes formatos que pode ser alvo de digitalização


azul.

SUSTENTABILIDADE

Há quase uma década que "gente bonita come fruta feia"

Projecto que vende a preços baixos cabazes de fruta da época e legumes que agricultores não conseguem vender a supermercados diz já ter evitado o desperdício de 3685 toneladas de alimentos. Estivemos num dos espaços da Fruta Feia, no Bonfim (Porto).

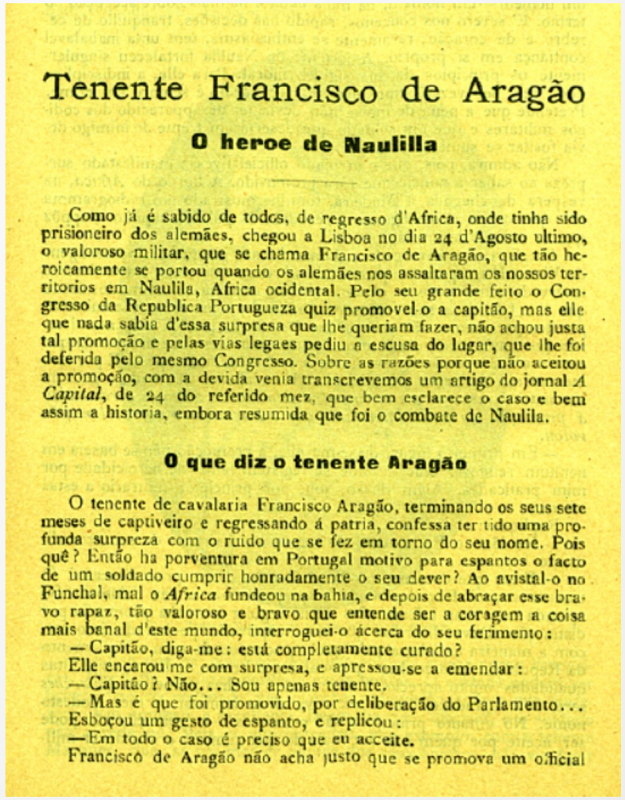
NELSON GARRIDO

Começou em Novembro de 2013 como um pequeno projecto cooperativo interessado em combater o desperdício alimentar e, no último dia 20 de Abril, abriu, no Mercado de São Domingos de Benfica, em Lisboa, o seu 15.º posto de venda de fruta e produtos hortícolas com um aspecto "imperfeito". Mais de oito anos e alguns prémios europeus depois, a Fruta Feia diz já ter evitado o desperdício de 3685 toneladas de alimentos, ajudando agricultores espalhados pelo país a escoar as peças de fruta e os legumes que, por não terem o aspecto desejado, não conseguem vender aos supermercados.

Criada por Isabel Soares, a iniciativa arrancou quase um ano depois de o Projecto de Estudo e Reflexão sobre o Desperdício Alimentar (ou, simplesmente, PERDA) ter revelado que Portugal desperdiçava um milhão de toneladas de alimentos por ano, número equivalente a 17% daquilo que era produzido a nível nacional (apesar de elevada, a percentagem estava, ainda assim, abaixo da média europeia). Querendo fazer a ponte entre agricultores e consumidores — para remover da equação a figura do distribuidor e, assim, poder praticar preços mais baixos —, a cooperativa começou a vender, a pessoas que dela se tornassem sócios, cabazes com fruta da época e produtos que, por serem demasiado disformes ou "feios", acabariam, de outra forma, no

G7

G8


Tenente Francisco de Aragão

O heroe de Naulila

Como já é sabido de todos, de regresso d'Africa, onde tinha sido prisioneiro dos alemães, chegou a Lisboa no dia 24 d'Agosto ultimo, o valoroso militar, que se chama Francisco de Aragão, que tão heroicamente se portou quando os alemães nos assaltaram os nossos territorios em Naulila, Africa ocidental. Pelo seu grande feito o Congresso da Republica Portugueza quiz promovel-o a capitão, mas elle que nada sabia d'essa surpresa que lhe queriam fazer, não achou justa tal promoção e pelas vias legaes pediu a escusa do lugar, que lhe foi deferida pelo mesmo Congresso. Sobre as razões porque não aceitou a promoção, com a devida venia transcrevemos um artigo do jornal *A Capital*, de 24 do referido mez, que bem esclarece o caso e bem assim a historia, embora resumida que foi o combate de Naulila.

O que diz o tenente Aragão

O tenente de cavalaria Francisco Aragão, terminando os seus sete meses de captiveiro e regressando á patria, confessa ter tido uma profunda surpreza com o ruido que se fez em torno do seu nome. Pois quê? Então ha porventura em Portugal motivo para espantos o facto de um soldado cumprir honradamente o seu dever? Ao avistal-o no Funchal, mal o *Africa* fundeou na bahia, e depois de abraçar esse bravo rapaz, tão valoroso e bravo que entende ser a coragem a coisa mais banal d'este mundo, interroguei-o acerca do seu ferimento:

— Capitão, diga-me: está completamente curado?

Elle encarou-me com surpresa, e apressou-se a emendar:

— Capitão? Não... Sou apenas tenente.

— Mas é que foi promovido, por deliberação do Parlamento...

Esboçou um gesto de espanto, e replicou:

— Em todo o caso é preciso que eu acceite.

Francisco de Aragão não acha justo que se promova um official


CASA FERREIRINHA

ESTEVA

RED WINE • VIN ROUGE

13% vol.

13% alc/vol.

PRODUCED AND BOTTLED BY • PRODUIT ET MIS EN BOUTEILLE PAR SOGRAPE VINHOS, S.A. V. N. DE GAIA, PORTUGAL - PRODUCT OF • PRODUIT DU PORTUGAL

750 ml

COMPLEXIDADE / COMPLEXITY / COMPLEXITÉ

TANINOS / TANINS

CORPO / BODY / CORPS

FRUTA / FRUIT

RETURN FOR REFUND WHERE APPLICABLE. CONSIGNÉ LÀ OÙ LA LOI LE PRESCRIT.

ORIGINAL LEGACY WINES

SOGRAPE

www.sograpevinhos.com

5 601007 001332

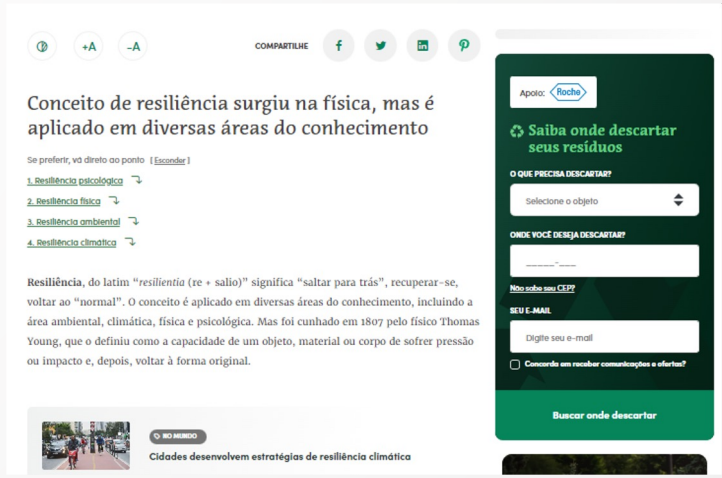
L5915601

G9

EXEMPLOS DE IMAGENS

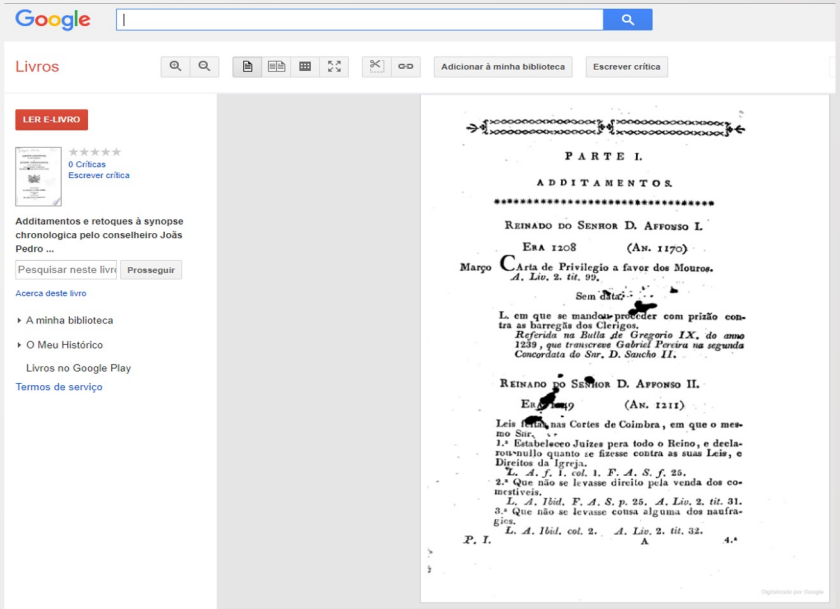
- Existe um conjunto diversificado de material em diferentes formatos que pode ser alvo de digitalização

G10



G11

G12



EXEMPLOS DE IMAGENS DADAS EM AULA

Fontes das imagens recolhidas e apresentadas:

- G1 - https://purl.pt/14355/1/obras/n48/n48_item59/P1.html
- G2 - https://tinyurl.com/ycxc7u7h
- G3 - https://cantigas.fcsh.unl.pt/cantiga.asp?cdcant=364&pv=sim
- G4 - https://amusearte.hypotheses.org/6492
- G5 - https://vdocuments.com.br/guia-de-compras-2016-revista-de-vinhos.html?page=1
- G6 - https://purl.pt/39776/2/
- G7 - https://tinyurl.com/2s3pc3ds
- G8 - https://tinyurl.com/y2baspze
- G9 - https://www.fgp.pt/store/esteva-tinto
- G10 - https://www.datocms-assets.com/33016/1664446617-gazela-vinho-verde-branco.pdf
- G11 - https://www.ecycle.com.br/resiliencia/
- G12 - https://tinyurl.com/5n6v2knn

Avaliação

PENSAR PARA O TRABALHO FINAL

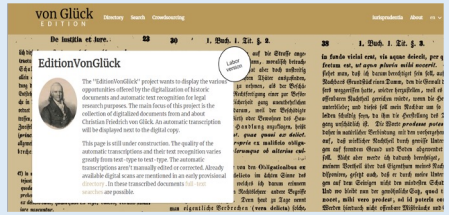
- O tópico?
- Quais documentos a digitalizar?
- Quantos documentos?
- Qual o melhor *software*?
- Esclarecer dúvidas sobre a digitalização …

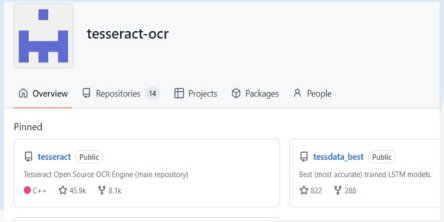
Obrigatório: Responder aos Inquéritos 1 e 2 (até dia 14/10/2022)

BIBLIOGRAFIA

Projetos

- https://rwi.app/iurisprudentia/en/vongl ueck
- https://rwi.app/iurisprudentia/en/v onglueck





Palestras online

Erasmus+

NOVAFCSH

EN	The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.
PT	O apoio da Comissão Europeia à produção desta publicação não constitui um aval do seu conteúdo, que reflete unicamente o ponto de vista dos autores, e a Comissão não pode ser considerada responsável por eventuais utilizações que possam ser feitas com as informações nela contidas.