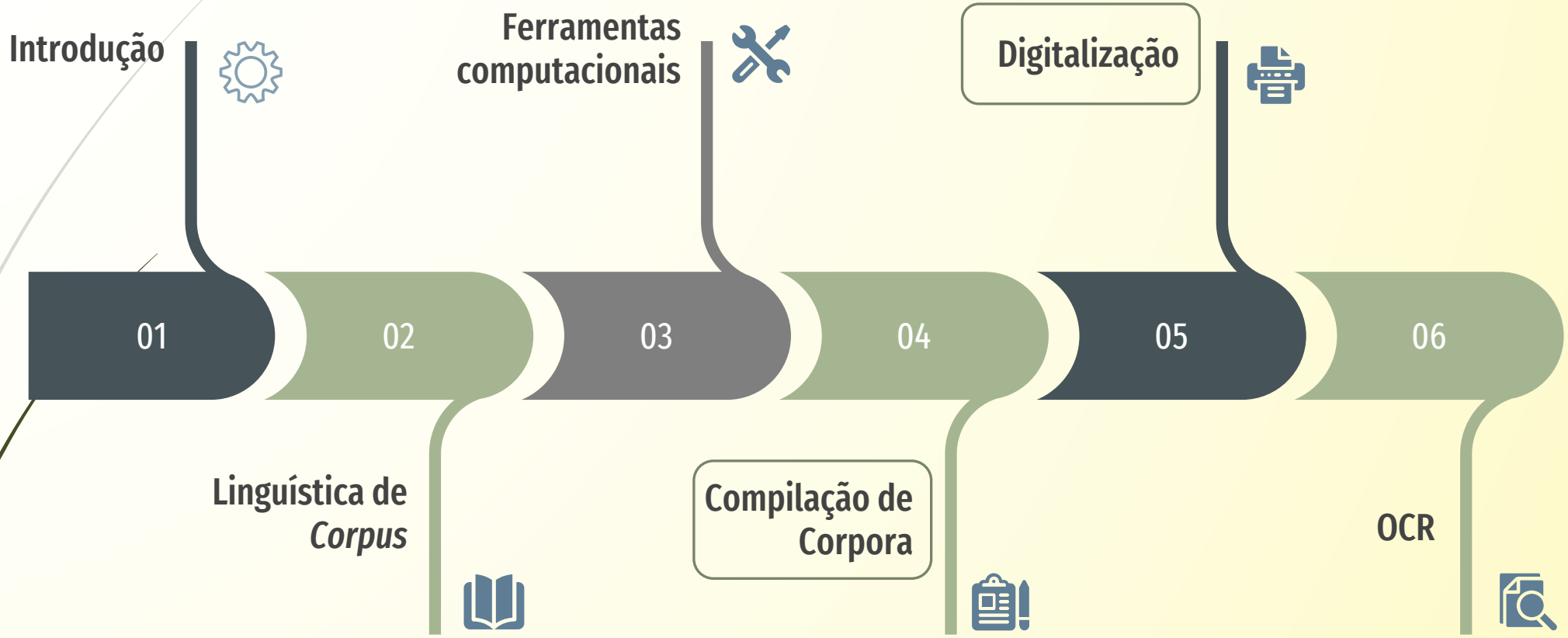
Humanities Going Digital (HUGOD)

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

Compilação de *Corpora* e *Digitalização*	Sessão 03	12/10/2022	Chiara Barbero Sílvia Barbosa

AULAS


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

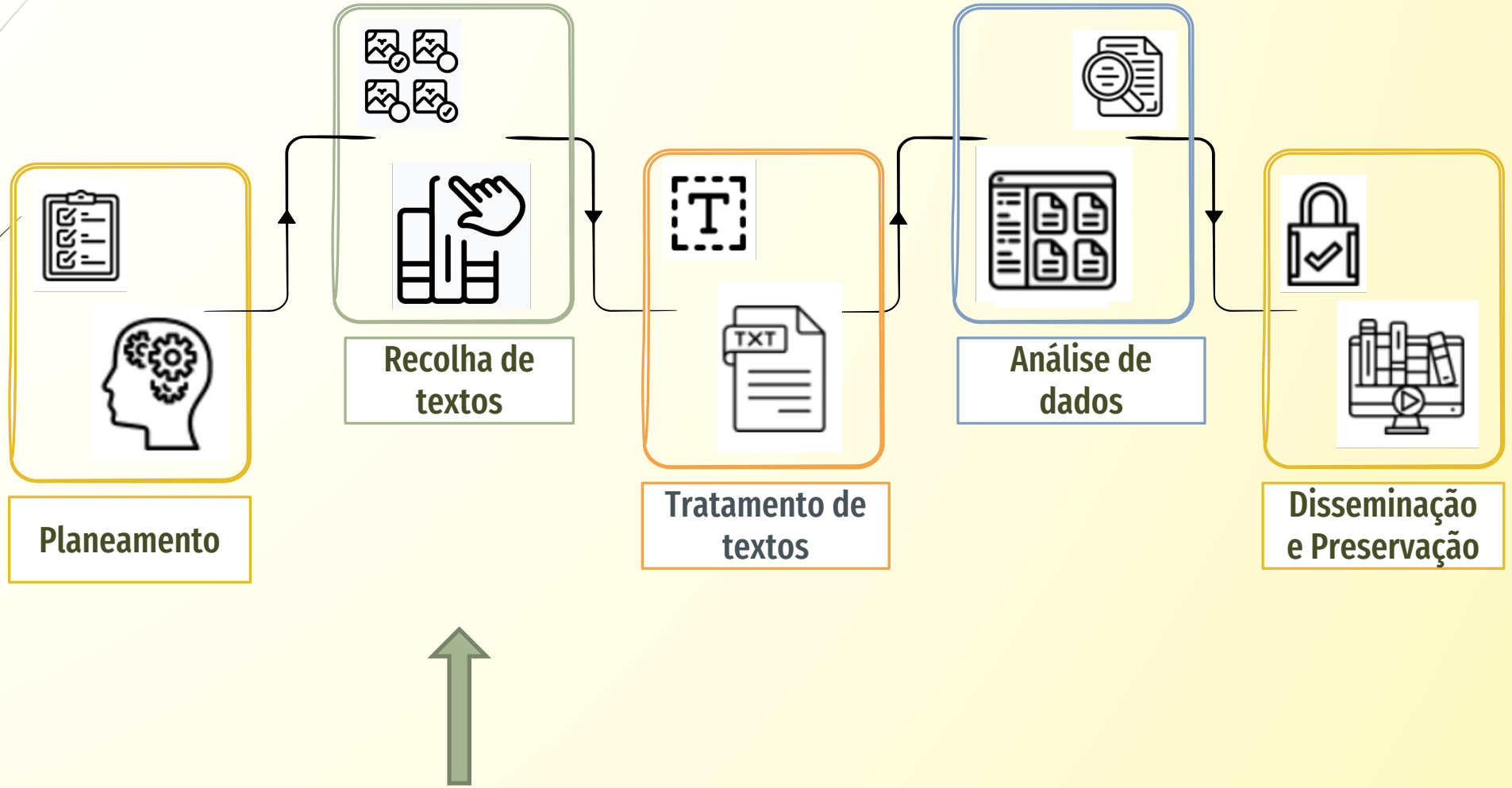
Procedimento de trabalho em Linguística de *Corpus*

Planeamento

- **Captura** de imagem (criação de réplica digital)
- **Tratamento** (conversão de uma imagem para um formato editável) → limpeza do texto, atribuição de metadados
- **Disseminação** (exploração do conteúdo digital obtido)
- **Preservação** digital a longo prazo (medidas que orientam no sentido de assegurar que o conteúdo digital pesquisável seja organizado em coleções que sejam mantidas acessíveis e atualizadas no futuro)

Planear é uma tarefa altamente importante, definir o objetivo do projeto de digitalização, identificar as prioridades, minimizar os riscos.

Procedimento de trabalho em Linguística de *Corpus*



Compilação de *Corpora*

Três possibilidades para a **Compilação de *Corpora***:

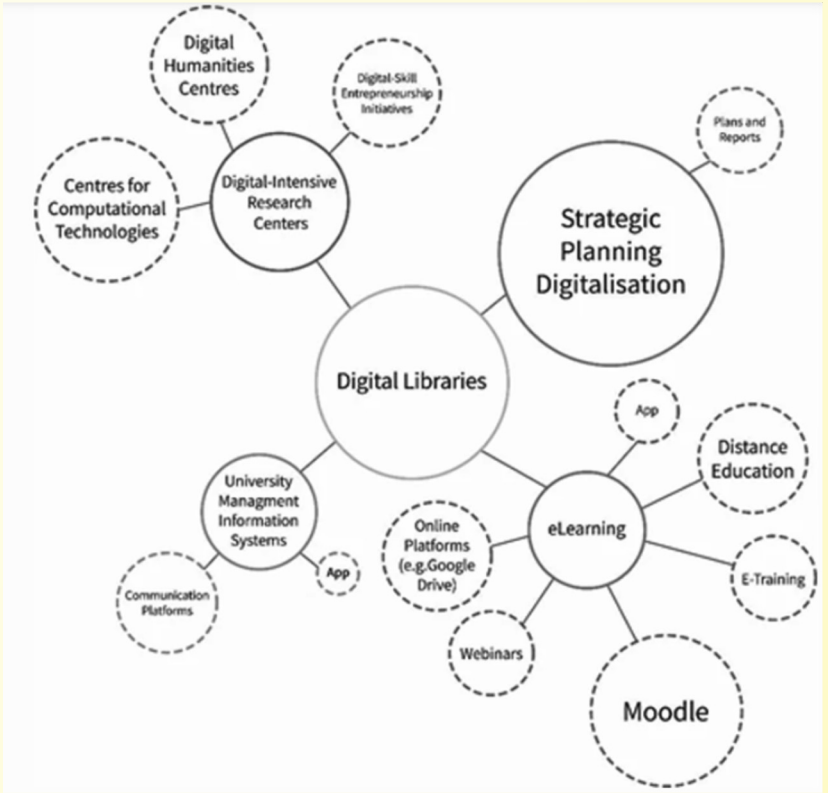
1. Digitalização
2. Consulta de materiais já existentes a partir de repositórios/base de dados
3. Ferramentas de recolha (ex. *Bootcat*)

Digitalização e HD

As HD permitem:

- observar o impacto da tecnologia na vida humana
- estudar as disciplinas das humanidades de uma forma diferente
- interligar melhor o passado, o presente e o futuro
- permitir uma maior e mais rápida análise e reflexão de aspetos anteriormente mais difíceis de captar

Digitalização é uma estratégia fundamental para apoiar as HD



https://link.springer.com/chapter/10.1007/978-3-030-56316-5_34

Digitalizar: porquê?

Bibliotecas, acervos, repositórios, coleções

não é suficiente para registar e preservar património coletivo da humanidade

Biblioteca de Alexandria

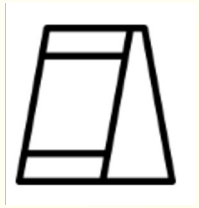
acervo ardeu e perdeu-se para sempre obras únicas da humanidade

Science History Images

acidente (incêndio, inundação), catástrofes naturais (cheias, tsunami, …), pilhagens, roubos, guerras, …

- permitir a preservação.
- permitir a acessibilidade.

Do formato papel até ao formato digital



vestígio de um momento
social – cultural - político - histórico

Do formato papel até ao formato digital

Vantagens da digitalização

- Facilita o acesso remoto e imediato a um conjunto de dados;
- Permite encontrar coleções/obras que estavam fora de circulação ou dispersas;
- Auxilia na preservação de material sensível/frágil;
- Possibilita a inclusão de diferentes formatos (tamanhos e resoluções);
- Proporciona cópia de obras a custo reduzido.

(https://library.princeton.edu/digital-collections/optimized-ocr)

Problemas da digitalização

- Problema de financiamento;
- questão de direitos de autor;
- questões de manuseamento de objetos antigos e frágeis,
- custos no processo (digitalizar, armazenar).

Pessoas e/ou instituições *não podem, nem conseguem* digitalizar tudo o que têm

um plano bem elaborado para a seleção do material.

Exemplos de projetos nacionais

- Literatura - https://bndigital.bnportugal.gov.pt/
- Literatura - https://cantigas.fcsh.unl.pt/manuscritos.asp
- Arte - https://www.archivesportaleurope.net/pt
- História/ militar - https://ahu.dglab.gov.pt/
- História / política - https://www.parlamento.pt/Parlamento/Paginas/RecursosEletronicos.aspx
- Botânica - https://arquivodebotanica.uc.pt/index.php
- Outros ...

Exemplos de projetos internacionais

- https://www.europeana.eu/pt/collections
- https://www.digitaltreasures.eu/
- https://www.distant-reading.net/
- Outros …

Digitalização ⟹ criar uma réplica digital ⟹ de um objeto tradicional

15 Compilação de *Corpora*

Três possibilidades para a **Compilação de *Corpora***:

1. Digitalização
2. Consulta de materiais já existentes a partir de repositórios/base de dados
3. Ferramentas de recolha (ex. *Bootcat*)

Repositórios digitais: coleção de informação digital

"a digital repository is where digital content, assets, are stored and can be searched and retrieved for later use. A repository supports mechanisms to import, export, identify, store and retrieve digital assets. (..) Digital repositories may include research outputs and journal articles, theses, e-learning objects and teaching materials or research data."

(https://www.yumpu.com/en/document/read/15608201/digital-repositories-jisc)

Tipos de repositórios:

- Temáticos: reúnem conteúdos de disciplinas ou assuntos específicos (arXiv; PubMed Central)
- Institucionais: repositórios criados por instituições de investigação científica (RUN, Repositório Ulisboa)
- De dados científicos: criados por diferentes tipos de organizações (Rcaap)

(https://openscience.usdb.uminho.pt/?page_id=348)

Outros exemplos de plataformas e bases de dados onde encontrar material eletrónico

- Bibliotecas
- Revistas
- Universidades
- Centros de Investigação
- Repositórios
- Infraestruturas
- Projetos já desenvolvidos

- https://tel.archives-ouvertes.fr/
- https://www.openedition.org/
- http://www.bnportugal.gov.pt/
- https://rossio.fcsh.unl.pt/
- https://zenodo.org/
- https://pangaea.de
- https://www.biodata.pt/
- https://www.dariah.eu/
- https://www.clarin.eu/
- https://www.gotriple.eu/
- https://www.cienciavitae.pt/
- https://orcid.org/
- https://www.researchgate.net/
- https://www.academia.edu/
- …

Compilação de *Corpora*

Três possibilidades para a **Compilação de *Corpora***:

1. Digitalização
2. Consulta de materiais já existentes a partir de repositórios/base de dados
3. Ferramentas de recolha (ex. *Bootcat*)

19 Compilação de *Corpora*

Ferramentas para a extração e compilação automática de *corpora* a partir da web, tendo como input "termos-sementes", i.e. termos que sejam expectáveis de serem típicos/representativos no âmbito do domínio de interesse.

Software:

- MonkeyLearn
- Google Cloud NLP
- IBM Watson
- Amazon Comprehend
- AYLIEN
- Thematic
- MeaningCloud

20 Compilação de *Corpora*

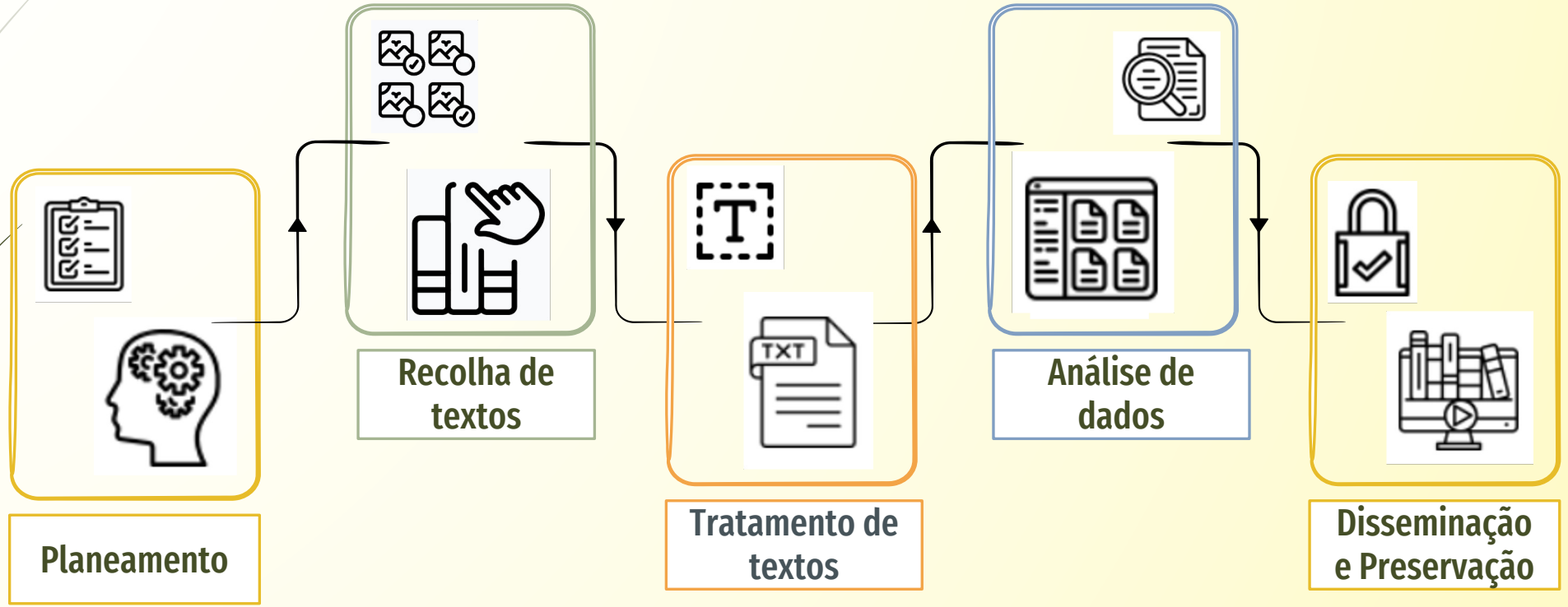
- seleção de URL relevantes (de acordo com os critérios estabelecidos)
- os termos-sementes são combinados de forma aleatória e cada combinação é usada como uma *string* de consulta num motor de busca (Google, Bing, Yahoo etc.)
- é estabelecido um intervalo de número de resultados (páginas) consideradas
- cada resultado extraído (*download*) é formatado como texto (geralmente as ferramentas são treinadas para eliminar automaticamente publicidades, menus de navegação e outros conteúdos linguisticamente irrelevantes)
- compilação da coletânea de textos

(https://marcobaroni.org/publications/lrec2004/bootcat_lrec_2004.pdf)

(https://www.sketchengine.eu/wp-content/uploads/2015/03/WebBootCaT_web_tool_2006.pdf)

Procedimento de trabalho em Linguística de *Corpus*

Procedimento de trabalho em Linguística de *Corpus*



Disseminação e Preservação

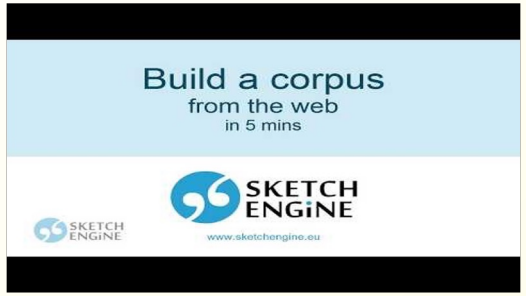
Preservação digital a longo prazo (medidas que orientam no sentido de assegurar que o conteúdo digital pesquisável seja organizado em coleções que sejam mantidas acessíveis e atualizadas no futuro)

- Disseminação
  - criando um site/aplicação
  - através de um ficheiro TXT ou XML
- Manutenção
  - Disponibilizar o material não é suficiente deve ser acautelada a manutenção (arquivo, para corrigir problemas, responder a utilizadores) – situações de projetos/instituições

Bibliografia Introdutória

Palestras online

- https://www.youtube.com/watch?v=VjHC4lMop-s



Livros e artigos

- http://www.digitalhumanities.org/dhq/vol/8/4/000196/000196.html
- https://www.digitisation.eu/about/
- https://docs.sslmit.unibo.it/doku.php?id=bootcat:tutorials:basic_1
- https://www.sketchengine.eu/wp-content/uploads/2015/03/WebBootCaT_web_tool_2006.pdf

Trabalho em aula

- Organizar os grupos de trabalho (2 ou 3 pessoas)
- Decidir o tópico

Para pensar …

- **O quê?**
- Porquê?
- Como?
- Para quem?
- **Qual o domínio / subdomínio?**
- Qual a Língua de trabalho?
- Qual o Enquadramento teórico?
- Qual (quais) o Objetivo(s)

Obrigatório: Inserir info. no MOODLE (Fórum Notícias)

Responder ao Inquérito 1

Avaliação

Trabalho em aula

- Inserir info no MOODLE (Fórum)
- Ler a bibliografia disponibilizada

Para a próxima aula

- Explorar os repositórios/bases de dados/plataformas sugeridos
- Escolher e digitalizar os documentos para serem trabalhados próxima aula

FIM DA SESSÃO

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.