

CURADORIA DE DADOS EM REPOSITÓRIOS: POLÍTICAS, PRÁTICAS E CASOS DE USO

Grupo de trabalho do Fórum GDI – Repositórios de Dados

Pedro Principe, Rita Gonçalves, André Vieira, Joaquim Santos, Sara Almeida e Margarida Alves

25 de novembro de 2022, Universidade de Évora

AGENDA



9.º FÓRUM
GESTÃO DE DADOS
DE INVESTIGAÇÃO

24 e 25 DE NOVEMBRO 2022
UNIVERSIDADE DE ÉVORA

- **Curadoria em repositórios de dados de investigação**
 - As tarefas essenciais na curadoria de dados: CURATION checklist,
 - *Pedro Príncipe (Universidade do Minho) e Rita Gonçalves (Universidade de Aveiro)*
 - Preocupações e problemas comuns na curadoria de (meta)dados,
 - *André Vieira (Universidade do Minho)*
- **Casos de uso de curadoria no ciclo de vida dos dados**
 - Dados de Biodiversidade: especificidades da curadoria de dados e a publicação no GBIF,
 - *Joaquim Santos (Universidade de Coimbra – Herbário)*
 - Dados Oceanográficos: boas práticas,
 - *Sara Almeida e Margarida Alves (Instituto Hidrográfico)*
- **Lacunas e pontos críticos no trabalho de curadoria de dados**



1. Curadoria em repositórios de dados de investigação

As tarefas essenciais na curadoria de dados: CURATION checklist
Preocupações e problemas comuns na curadoria de (meta)dados

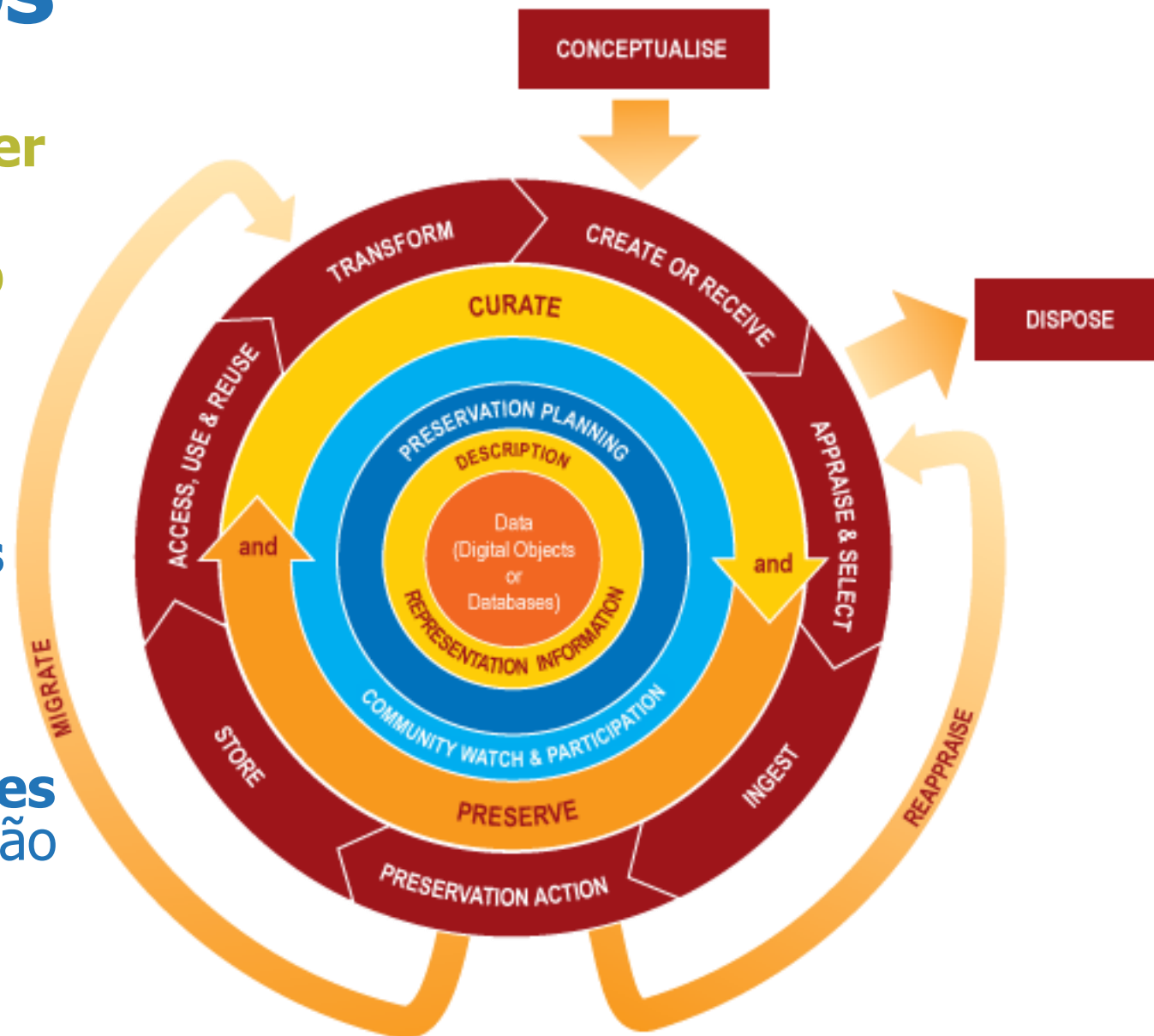
Curadoria de Dados

Modelo do Digital Curation Center

Visão geral gráfica das etapas para uma gestão, curadoria e preservação bem sucedida dos dados.

Antes de mais:

- A curadoria ocorre em **diferentes etapas** do ciclo de vida da investigação e dos dados.
- É preciso assegurar diferentes tarefas de curadoria para **diferentes propósitos** (publicação, preservação



CRIAR OS DADOS: conceber a investigação, planos de gestão de dados, localizar datasets existentes, obter consentimentos, recolher e gerir dados, capturar e criar metadados.

REUTILIZAR OS DADOS: Ajudar a encontrar fontes de dados, catálogos de dados, repositórios de dados, aconselhamento à citação.

DAR ACESSO AOS DADOS: Repositórios de dados, registo de PIDs, apoiar a definição de licenças, publicação (dados), VREs, armazenamento.

Planear e
criar
dados

PROCESSAR OS DADOS: Apoiar a criação de dados, ferramentas de documentação, organização e armazenamento, lab notebooks, conselhos sobre metadata standards.

ANALISAR OS DADOS: Ferramentas de Workflows e versionamento, serviços de transcrição, software de análise de dados, apoio estatístico, limpeza e anonimização de dados.

PRESERVAR OS DADOS: Arquivo de dados, editar/criar metadados, repositórios de dados, enriquecer a documentação, serviços de preservação (armazenamento/backups).

Reutilizar
os dados

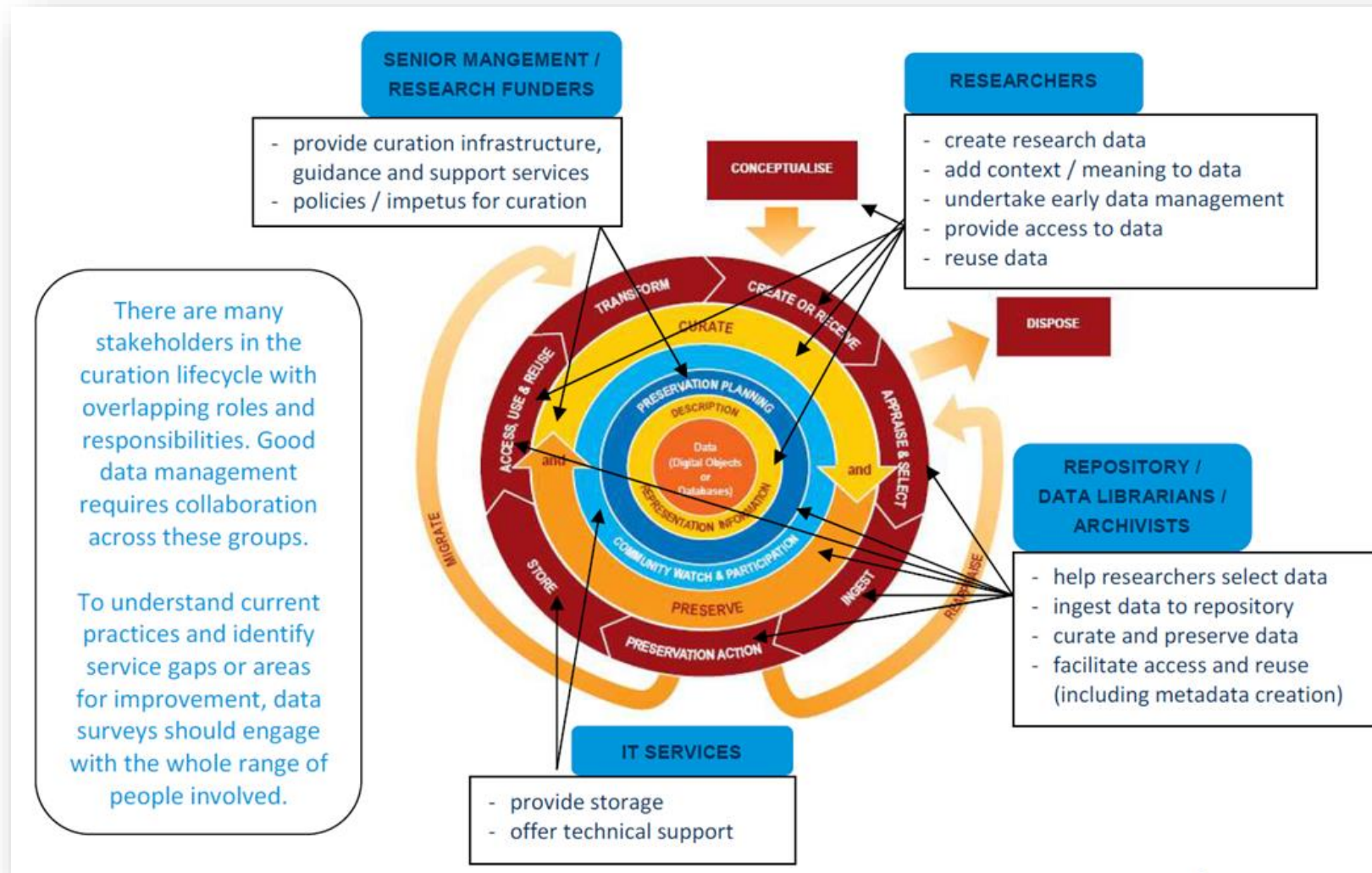
Processar
os dados

Analisar
os dados

Preservar
os dados

Dar
acesso
aos dados

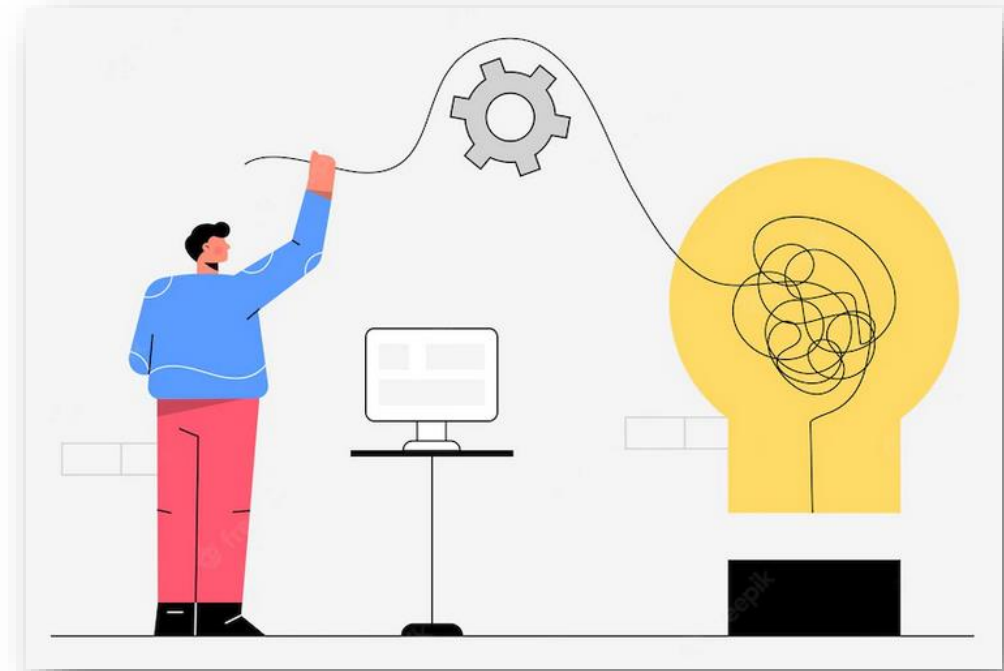
Curadoria em diferentes fases da gestão de dados e com vários papéis e responsabilidades



Curadoria em REPOSITÓRIOS de dados de investigação

Premissas

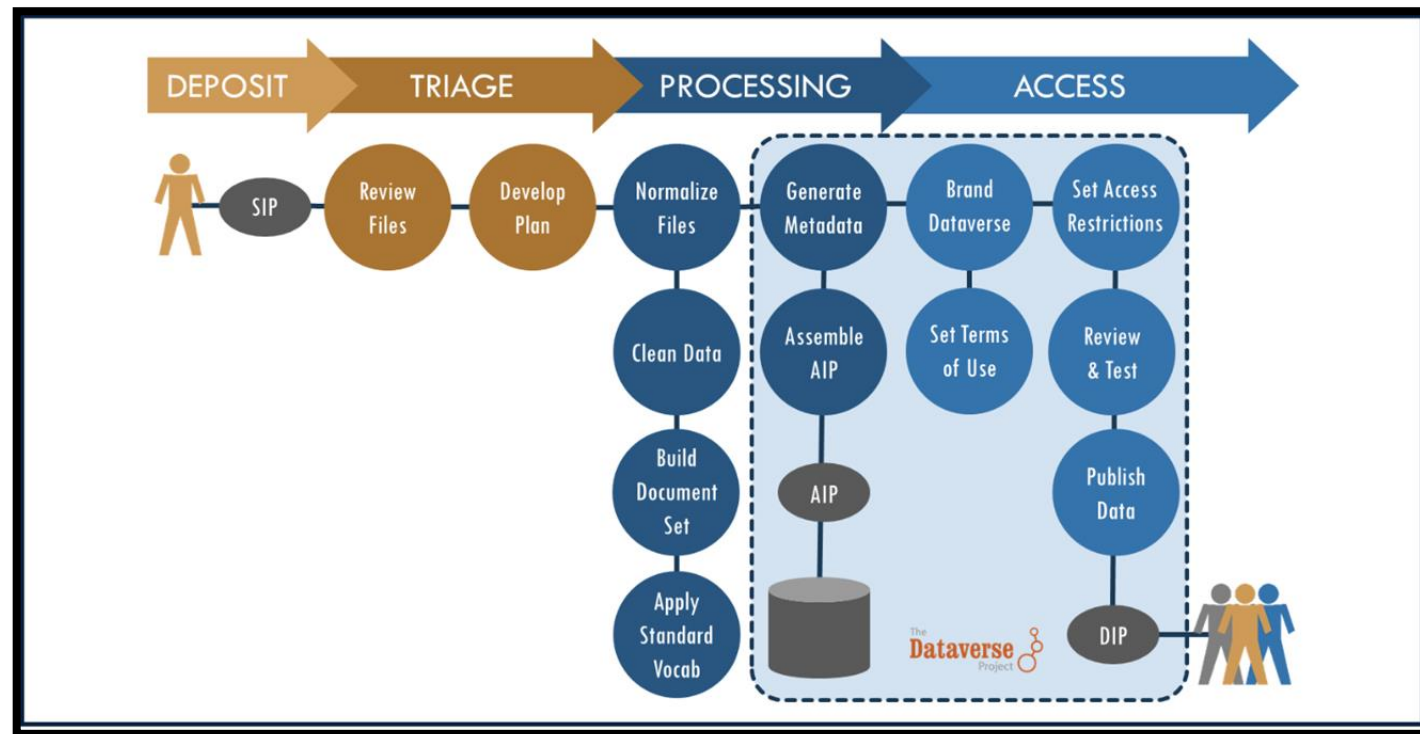
- Não existe um modelo único para a curadoria de dados.
- Depende de condições e recursos locais, políticas, prioridades e estratégia institucional.
- Resulta das ferramentas, processo e instrumentos usado.



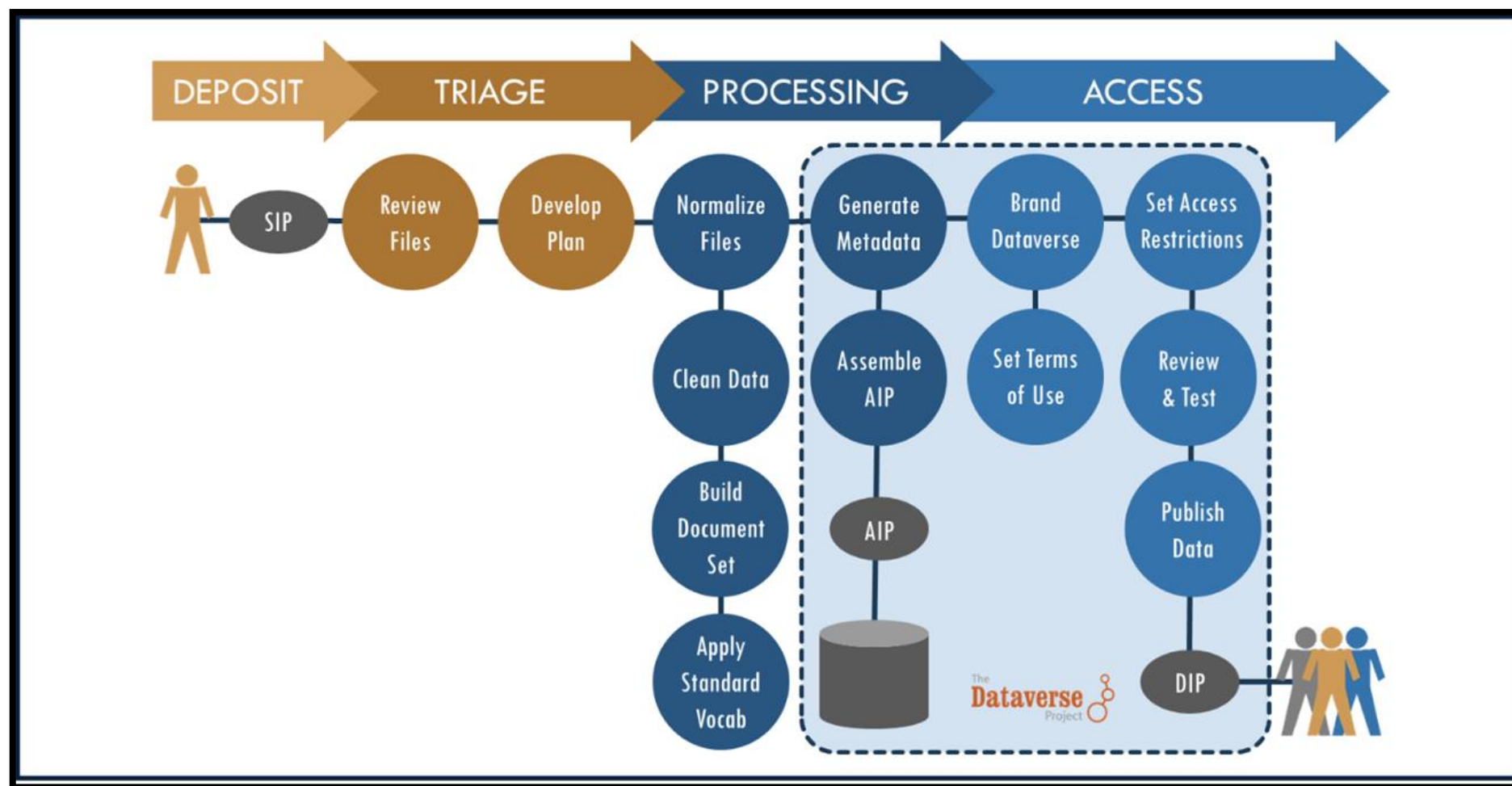
Fluxo de trabalho da curadoria de dados

Quatro fases primárias:

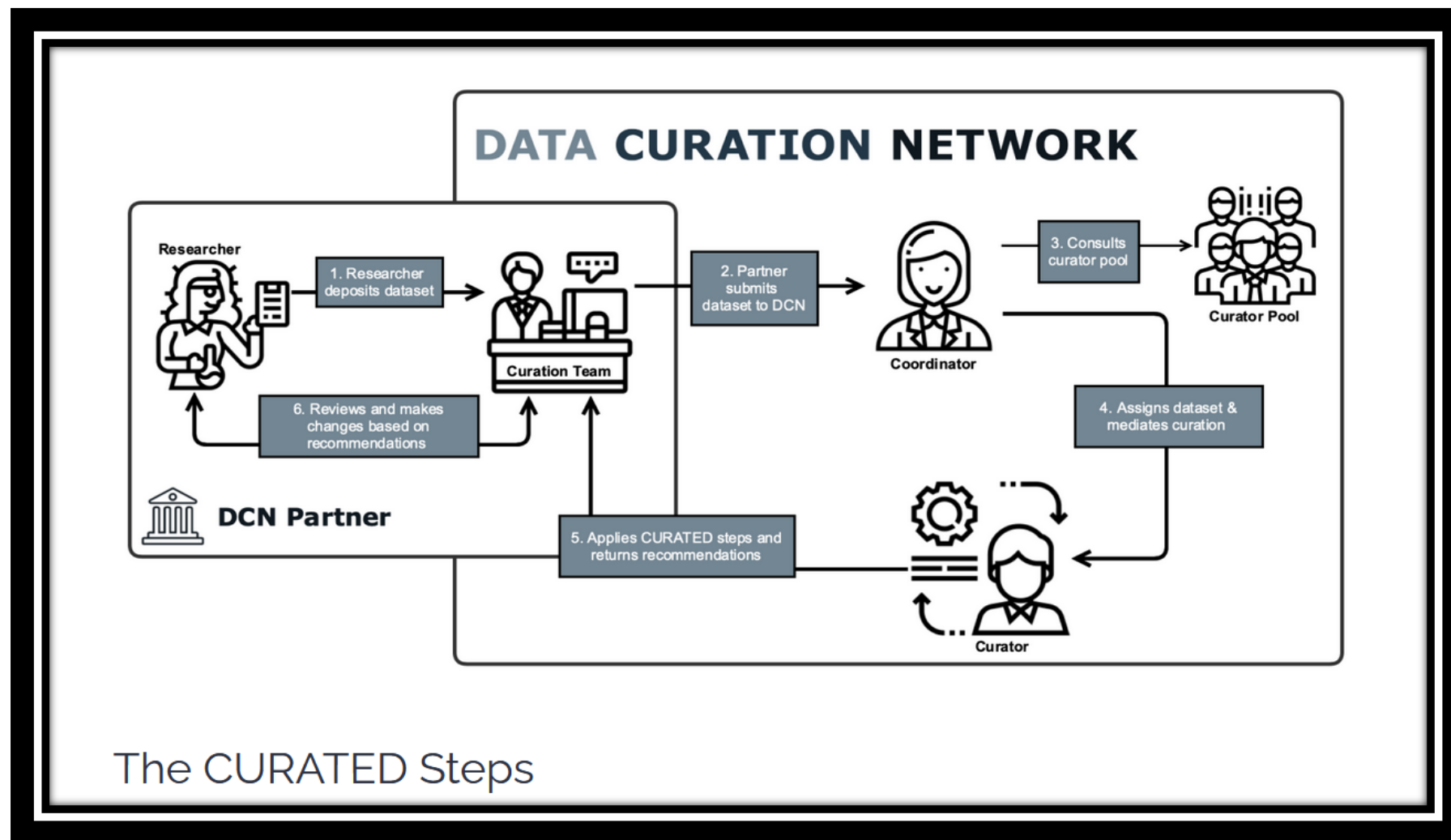
1. Depósito
2. Seleção
3. Processamento
4. Acesso



Odum Institute Data Archive data curation workflow



The DCN Curation Workflow



Data Curation Network

Data Curation Primers

Data curation primers are peer-reviewed, living documents that detail a specific subject, disciplinary area or curation task and that can be used as a reference to curate research data. The links below are to the live, editable copies in [GitHub](#) while archived pdf versions are available in the [DCN repository](#).

Authors of primers include archivists and data librarians who attended the 2018-2020 Specialized Data Curation Workshops presented by the Data Curation Network (IMLS RE-85-18-0040-18).

Acrobat PDF Primer

Creators: Peace Ossom-Williamson, Nicole Contaxis, Margaret Lam and Adam Kriesberg
Mentor: Jake Carlson

ATLAS.ti Primer

Creator: Margarita Corral
Affiliated contributor: Hannah Hadley
Mentor: Dave Fearon

Confocal Microscopy Image Primer

Creators: Susan Ivey, Amy Koschoffer, Gretchen Sneff and Huajin Wang
Mentor: Lisa Johnston

Consent Forms Primer

Creators: Shanda Hunt, Alicia Hofelich Mohr and Rachel Woodbrook

Databases Primer

Creator: Xuying Xin
Mentor: Dave Fearon

Atividades de Curadoria específicas ao nível disciplinar e em sistemas e instrumentos próprios

GeoJSON Primer

Mentor: Mara Blake

Creators: Nadia Dixson, Genevieve Milliken, Keshav Mukunda, Reina Murray and Rachel Starry

Mentor: Mara Blake

GeoTIFF Primer

Creators: Courtney Kearney, Nick Ruhs, Mara Sedlins, Tracy Tien, Jessica Trelogan and John Watts

Mentor: Jennifer Moore

Google Docs Primer

Creators: Nadia Dixson and Adam Kriesberg
Mentor: Sophia Lafferty-Hess

Human Participants Data Essentials Primer

Creators: Jen Darragh, Alicia Hofelich Mohr, Shanda Hunt, Rachel Woodbrook, Dave Fearon, Jennifer Moore and Hannah Hadley

ISO Images Primer

Creators: Kate Barron and Jonathan Bohan
Mentor: Cynthia Hudson Vitale

Java Primer

Creators: Daina Bouquin, Sophie Hou, Matthew Rios and Lisa Woodbrook

Matlab Primer

Creators: Sam Scobie and Susan Borda

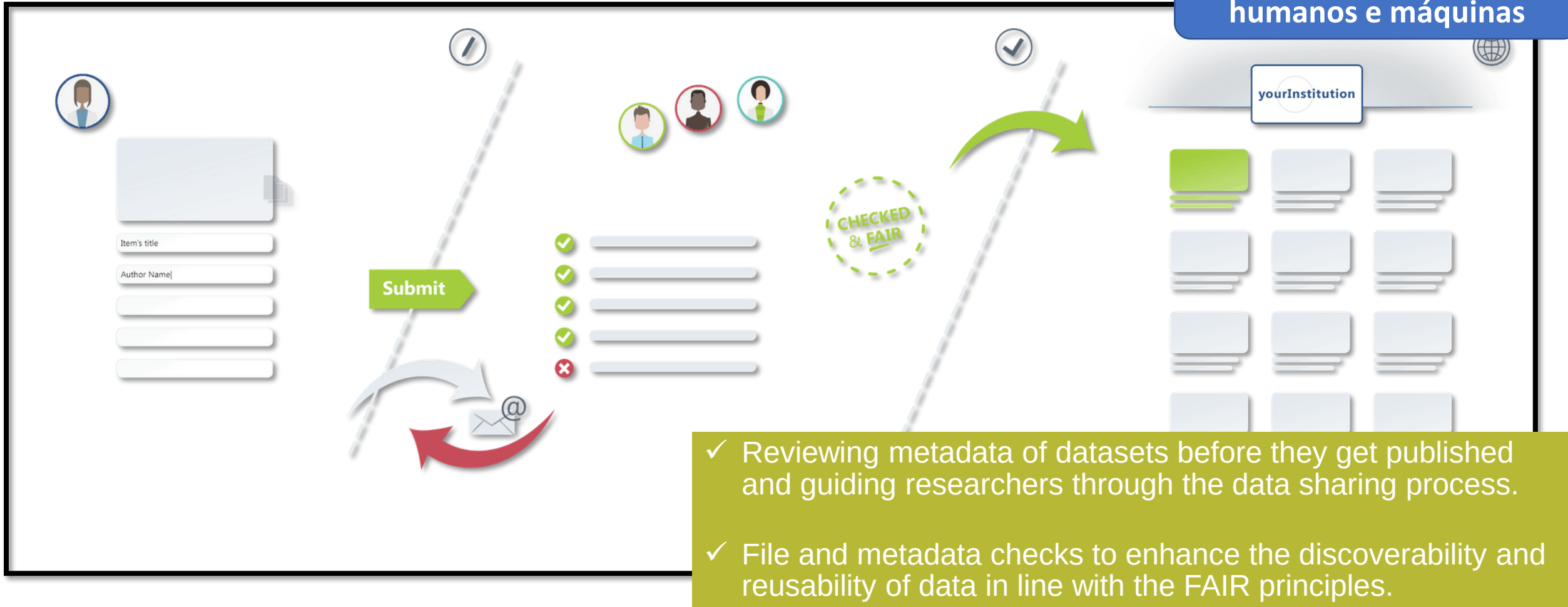
Microsoft Access Primer

Creators: Fernando Rios
Mentor: Dave Fearon

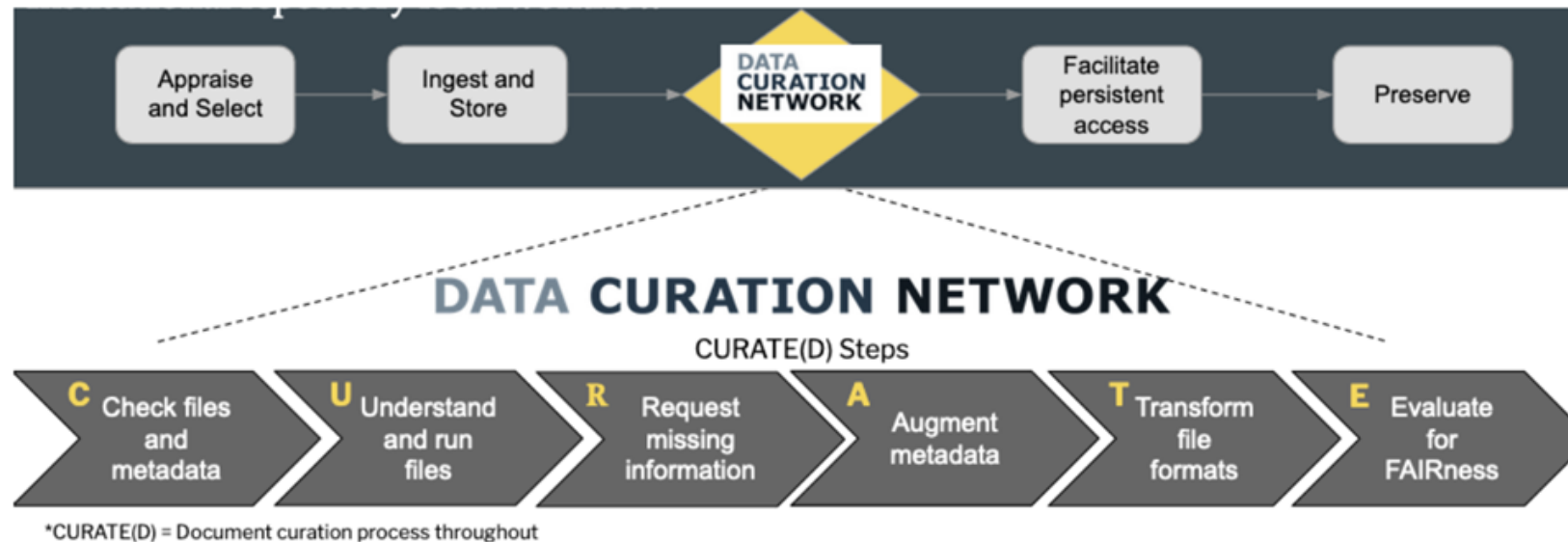
<https://datacurationnetwork.org/outputs/data-curation-primers/>

Figshare Curation Service

Curadoria especializada:
humanos e máquinas



DCN – Curation steps and checklist



CC-BY Suggested Citation: Data Curation Network (2022). "CURATE(D) Steps and Checklist for Data Curation. version 2" <http://z.umn.edu/curate>. (also found at <https://datacurationnetwork.org>)

<https://docs.google.com/document/d/1RWt2obXOOeJRRFmVo9VAkl4h41cL33Zm5YYny3hbPZ8/edit#>

Tradução e adaptação do

Dataverse Curation Guide

Dataverse Curation Guide Working Group

New Digital Research Infrastructure Organization (NDRIO) / Portage

Três níveis de curadoria

Nível 1

As etapas mínimas necessárias para publicar com êxito num repositório e tornar o conjunto de dados localizável.

Nível 2

Atividades que potenciam a descoberta dos conjuntos de dados e ajudam a garantir a sua usabilidade ao longo do tempo.

Nível 3

Ações intensivas de curadoria destinadas a preparar os conjuntos de dados para preservação e melhorar as hipóteses dos dados e código serem utilizados para reproduzir ou replicar um estudo associado.

Guia 'CURATION'

C

Check

Assegurar que todos os dados e metadados necessários para publicar o conjunto de dados com sucesso estão reunidos e em condições de publicação.

U

Understand

Garantir que o conjunto de dados está bem descrito e que os utilizadores terão uma visão clara a que se referem os dados e como podem ser utilizados.

R

Recomend

Solicitar informações adicionais ao depositante ou sugerir alterações aos metadados e ficheiros de forma a potenciar a descoberta e a utilização dos dados de acordo com os princípios FAIR.

A

Augment

Melhorar a submissão para facilitar a capacidade de descoberta e uso.

T

Transform

Assegurar que o conjunto de dados utiliza o maior número possível de formatos abertos e de uso generalizado.

I

Include

Facilitar a reutilização, bem como a correta atribuição e creditação dos dados, incluindo os identificadores persistentes relevantes e as informações de licenciamento adequadas.

O

Optimize

Avaliar a adequação do conjunto de dados aos princípios FAIR e tomar medidas para melhorar a descoberta, acessibilidade, interoperabilidade e reutilização dos dados.

N

Note Down

Certificar-se de que se faz um registo exato do trabalho de curadoria.

Estrutura do guia de curadoria

C	U	R	A	T	I	O	N
Check	nderstand	ecommend	ugment	ransform	nclude	ptimize	ote Down
Nível 1	Nível 2 Nível 3	Nível 1	Nível 2	Nível 3	Nível 2 Nível 3	Nível 3	Nível 2

Sim	Não	Levanta questões	N/A	A documentação inclui...
☐	☐	☐	☐	Informação de contexto sobre os dados (como os dados foram recolhidos ou produzidos, o objetivo da investigação).
☐	☐	☐	☐	Descrição das convenções de nomenclatura dos ficheiros e da estrutura dos ficheiros, se importante.
☐	☐	☐	☐	Registo de como os dados foram modificados ou processados.
☐	☐	☐	☐	Informações sobre confidencialidade e quaisquer restrições impostas ao uso dos dados.
☐	☐	☐	☐	Nomes de descritores e variáveis, informações sobre valores permitidos e unidades de medida, códigos e classificações, se aplicável.
☐	☐	☐	☐	Descrição do ambiente computacional necessário para executar qualquer código incluído (sistema operacional, pacotes de software e dependências).

Sim	Não	Levanta questões	N/A	Identifique os formatos e o software usado na criação de ficheiros...
☐	☐	☐	☐	O conjunto de dados contém ficheiros com formatos proprietários ou inapropriados para preservação? Para orientações, consulte os formatos de ficheiros elencados em DataverseNO .
☐	☐	☐	☐	Os ficheiros proprietários podem ser transformados em formatos não-prorietários sem que ocorra perda de informação?
☐	☐	☐	☐	Caso os ficheiros não possam ser transformados, a sua documentação descreve: - O formato e o conteúdo dos ficheiros? - O software ou instrumento que gerou os ficheiros? - O software necessário para apresentar ou manipular os ficheiros? - Uma alternativa freeware para aceder aos ficheiros (se disponível)?
☐	☐	☐	☐	A documentação descreve os dados, as fórmulas usadas, os cabeçalhos de colunas, os nomes de variáveis, etc. por forma a que os dados sejam usáveis.
☐	☐	☐	☐	A documentação está num formato adequado (ex.: PDF/A ou .txt).

Sim	Não	N/A	
☐	☐	☐	Um ficheiro de registo foi criado para documentar a seguinte informação: ☐ DOI, identificador interno do item ou outra forma de relacionar o ficheiro de registo com o conjunto de dados ☐ Identificar problemas descobertos no decorrer do processo de curadoria ☐ Questões e recomendações de alterações significativas para o investigador ☐ Resumo detalhado das respostas do investigador às sugestões e questões ☐ Qualquer alteração efetuada no conjunto de dados durante o processo de curadoria, incluindo: ☐ Alterações aos ficheiros existentes e documentação ☐ Nome de ficheiros adicionados ou removidos do conjunto de dados e qual o motivo ☐ Listagem dos termos alterados, adicionados ou removidos dos <u>metadados</u>

Guia disponível em:

<https://forumgdi.rcaap.pt/grupos-de-trabalho/gt-repositorios-de-dados/>

Vamos navegar agora por algumas tarefas essenciais para verificação de curadoria.

Boletim de Curadoria de dados em repositórios

<https://hdl.handle.net/1822/78558>



data RepositórioUM

Guia Informativo 2022
<https://datarepositorium.uminho.pt>

Para Investigadores, Docentes e Responsáveis de Unidades de Investigação
(curadores de dados no DataRepositórioUM)

Curadoria de dados em repositórios

Guia de tarefas na publicação e curadoria de dados:

Nível 1. Ao assegurar este nível de curadoria para os conjuntos de dados a serem publicados no **Dataverse**, está a assumir o nível mínimo necessário para que estes possam ser localizados.

Nível 2. Com este nível de curadoria potencia a descoberta e ajuda a garantir a usabilidade dos conjuntos de dados ao longo do tempo.

Nível 3. Atingindo este nível de curadoria está a garantir a reprodutibilidade e preservação aos conjuntos de dados.

Letra	Definição	Principais verificações / tarefas	Nível 1	Nível 2	Nível 3
	Check Assegurar que todos os elementos necessários para publicar com sucesso o conjunto de dados estão reunidos e em condições de publicação	O conjunto de dados foi depositado no dataverse adequado	X		
		Todos os ficheiros descritos na documentação (nota metodológica, ficheiro README) foram incluídos no conjunto de dados	X		
		Os campos de metadados requeridos foram preenchidos	X		
		A documentação de apoio foi incluída (nota metodológica, ficheiro README)	X		
		O Investigador assegura que os dados estão livres de quaisquer restrições em matéria de licenciamento e propriedade intelectual	X		
		O Investigador confirmou que os dados não são sensíveis e estão livres de informações que permitam a identificação	X		
	Understand Assegurar que o conjunto de dados está bem descrito e que os utilizadores finais terão uma imagem clara do que são os dados e como podem ser utilizados	A documentação de apoio é minuciosa, precisa e completa	X		
		Os ficheiros abrem corretamente e o conteúdo aparece como esperado	X		
		Os ficheiros e pastas são nomeados e estruturados de forma apropriada	X		
		Em caso de depósito de conjuntos de código, assegurar que este está bem documentado e produz os resultados esperados			X
		O conjunto de dados submetidos / depositados não contém dados sensíveis			X
		O conjunto de dados submetido não contém dados ou código fonte de terceiros			X
	Recommend Solicitar informações adicionais ao depositante ou sugerir alterações nos metadados e ficheiros, facilitando a possibilidade de os localizar e utilizar de acordo com os princípios FAIR	Dar prioridade às suas recomendações numa lista, para determinar quais os pedidos críticos ou imediatos, e quais os pedidos de prioridade mais baixa	X		
		Fazer um pedido claro de informação ao depositante	X		

[Ver mais sobre Princípios FAIR aqui - <https://doi.org/10.1038/sdata.2016.18>]

Para assegurar as principais tarefas de curadoria em repositórios... PREPARAR DADOS É ESSENCIAL

Como preparar dados para depositar e publicar no repositório

<https://doi.org/10.5281/zenodo.7082551>

Ver também:

<https://site.uit.no/dataverseno/deposit/prepare/>



Como preparar dados para depositar e publicar no repositório

Dicas práticas para Investigadores, curadores de dados, gestores de laboratório e responsáveis de unidades de investigação

Objetivo: Dotar os investigadores, gestores de dados e públicos relevantes, de ferramentas que promovam o depósito de dados, cumprindo determinados procedimentos, indo de encontro às orientações emanadas pelos organismos financiadores (ex. UE) e instituições nacionais e internacionais especializadas na GDI, no contexto da Ciência Aberta e na implementação dos princípios FAIR.

Formato dos dados

No início de um projeto existe a necessidade de criação de um **Plano de Gestão de Dados**, requisito de muito projetos financiados (ex. FCT e CE), sendo que uma das primeiras decisões a tomar é definir quais os formatos de ficheiros a utilizar, uma vez que existem **formatos padrão** e **formatos proprietários**. Estes últimos podem acarretar consequências negativas, tornando os dados menos interoperáveis. O grande foco será sempre a sua **preservação**, o **acesso** continuado aos dados, a sua **interoperabilidade** e, tanto quanto possível, o cumprimento dos **princípios FAIR**. Não existem garantias se formatos de ficheiros proprietários usados diariamente existirão no futuro. Por exemplo, formatos de ficheiros da Microsoft (Word e Excel), podem estar em uso agora, mas isto não invalida a possibilidade de no futuro se tornarem obsoletos.

Os **formatos dos ficheiros** podem apresentar-se com ou sem perdas de informação: ou seja, se os dados são ou não comprimidos. A compressão de ficheiros tem como objetivo a redução do tamanho dos mesmos por intermédio da remoção de informação redundante. Atualmente, podemos considerar os formatos abaixo indicados como apropriados e aceitáveis para preservação a longo prazo, assegurando assim o acesso continuado.

Aspeto a reter

Devemos referir que tomar os **dados FAIR** é cada vez mais uma responsabilidade conjunta dos investigadores e dos repositórios. Neste contexto e por forma a otimizar o arquivo, disseminação e reutilização de dados, o desenvolvimento de metadados ricos é um requisito dos **princípios FAIR** (Findable, Accessible, Interoperable and Reusable).

Tipos de dados	Apropriado	Aceitável	Não apropriado
Dados tabulares com metadados extensos	.csv - .hdf5	.txt - .html - .tex - .por	
Dados tabulares com metadados mínimos	.csv - .tab - .ods - SQL	.xml if appropriate DTD - .xlsx	.xls - .xlsb
Dados textuais	.pdf - .txt - .odt - .odm - .tex - .md - .htm - .xml	.pptx - PDF with embedded forms - .rtf	.doc - .ppt
Código	.m - .R - .py - .jy - .studio - .rmd - NetCDF	.sdd	.mat - .rdata
Dados de imagem digital	.tif - .png - .svg - .jpeg	.jpg - .jp2 - .tif - .tiff - .pdf - GIF - BMP	.indd - .ait - .psd
Dados de áudio digital	.flac - .wav - .ogg	.mp3 - .mp4 - .aif	
Dados de vídeo digital	.mp4 - .m2 - .avi - .mkv	.txt - .html - .tex - .por	.wmv - .mov
Dados geoespaciais	NetCDF, tabular GIS Attribute data, shp - .shx - .dbf - .prj - .sbn - PostGIS - .tif - .tiff - GeoJSON	.ogm - .webm	
Dados vetoriais e matriciais	.dwg - .dxf - .x3d - .x3dv - .x3db - .pdf - .PDF3D		
Dados genéricos	.xml - .json - .rdf		

Fonte: MOOC "O essencial da gestão de dados de investigação" em <https://www.raa.edu.pt/pt/cursos/o-essencial-da-gestao-de-dados-de-investigacao/>

Curation | Check :: Verificar / Conferir

Letra	Definição	Principais verificações / tarefas	Nível 1	Nível 2	Nível 3
	Check Assegurar que todos os elementos necessários para publicar com sucesso o conjunto de dados estão reunidos e em condições de publicação	O conjunto de dados foi depositado no dataverse adequado	X		
		Todos os ficheiros descritos na documentação (nota metodológica, ficheiro README) foram incluídos no conjunto de dados	X		
		Os campos de metadados requeridos foram preenchidos	X		
		A documentação de apoio foi incluída (nota metodológica, ficheiro README)	X		
		O Investigador assegura que os dados estão livres de quaisquer restrições em matéria de licenciamento e propriedade intelectual	X		
		O Investigador confirmou que os dados não são sensíveis e estão livres de informações que permitam a identificação	X		

Possui documentação de apoio

Metadados preenchidos

Confirmação de termos de uso

Curation | Understand :: Compreender

U

Understand

Assegurar que o conjunto de dados está bem descrito e que os utilizadores finais terão uma imagem clara do que são os dados e como podem ser utilizados

A documentação de apoio é minuciosa, precisa e completa

X

Documentação de apoio precisa

Os ficheiros abrem corretamente e o conteúdo aparece como esperado

X

Ficheiros corretos

Os ficheiros e pastas são nomeados e estruturados de forma apropriada

X

Em caso de depósito de conjuntos de código, assegurar que este está bem documentado e produz os resultados esperados

X

O conjunto de dados submetidos / depositados não contém dados sensíveis

X

O conjunto de dados submetido não contém dados ou código fonte de terceiros

X



Curation | Recommend :: Recomendar

R	Recommend Solicitar informações adicionais ao depositante ou sugerir alterações nos metadados e ficheiros, facilitando a possibilidade de os localizar e utilizar de acordo com os princípios FAIR	Dar prioridade às suas recomendações numa lista, para determinar quais os pedidos críticos ou imediatos, e quais os pedidos de prioridade mais baixa	X		
		Fazer um pedido claro de informação ao depositante (Ver mais sobre Princípios FAIR aqui - https://doi.org/10.1038/sdata.2016.18)	X		

Verificações do
objeto FAIR



Curation | Check :: Aprefeiçoar



Augment

Melhorar a submissão para facilitar a descoberta e a usabilidade

As descrições dos conjuntos de dados são detalhadas, precisas e completas

Os links para publicações, conjuntos de dados e outros recursos relacionados estão incluídos

X

X

Descrições completas



Curation | transform :: Converter

Formatos para preservação				
T	Transform Assegurar que o conjunto de dados é disponibilizado em formato aberto e de uso generalizado	Os formatos dos ficheiros são abertos e devidamente documentados		X

Curation | Include :: Incluir

Identificadores				
	Include Facilitar a reutilização, atribuição e crédito adequados aos dados, através da inclusão de identificadores persistentes e informação de licenciamento apropriada	Incluir identificadores persistentes sempre que possível (ex. DOI ou handle)		X
		Rever o licenciamento e termos de utilização do conjunto de dados (ex. Licenças Creative Commons - https://creativecommons.org/)		X

Curation | Optimize :: Otimizar



Optimize

Avaliar a adequação dos conjuntos de dados aos princípios FAIR e tomar medidas para otimizar a capacidade de localização, acessibilidade, interoperabilidade e reutilização dos dados

Avaliar o conjunto de dados e otimizar a adequação aos princípios FAIR

(Consultar esta ferramenta para avaliação do cumprimento dos Princípios FAIR - <https://fairaware.dans.knaw.nl/>)

FAIRification

X

Curation | Note down :: Registrar

Usabilidade				
N	Note Down Assegurar que fez um registo exato e documentado do seu trabalho de curadoria	Criar um registo de curadoria para documentar as suas decisões e ações		X



2. Casos de uso de curadoria no ciclo de vida dos dados

Dados de Biodiversidade: especificidades da curadoria de dados e a publicação no GBIF

Dados Oceanográficos: boas práticas

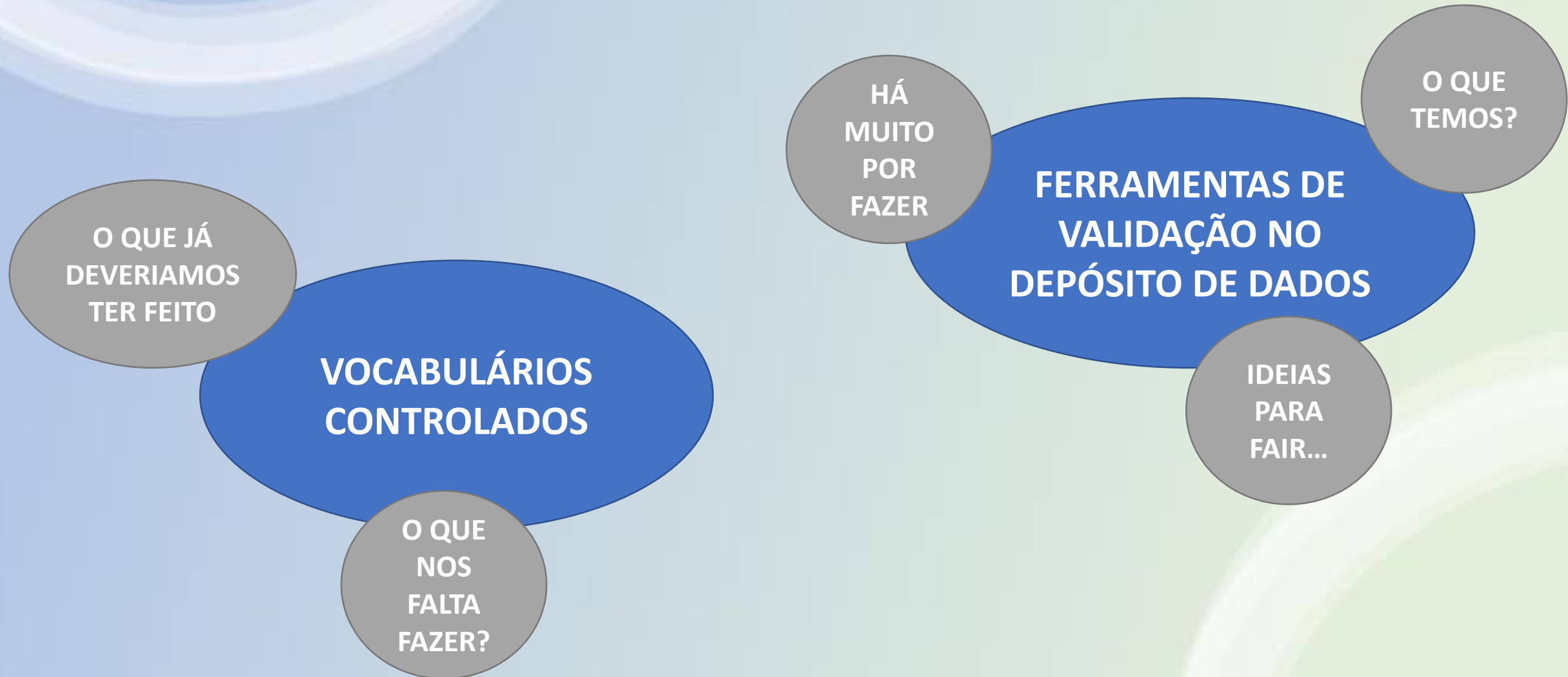


3. Lacunas e pontos críticos

no trabalho de curadoria de dados

DEBATE

Lacunas e pontos críticos



Lacunas e pontos críticos

Mais...





FÓRUM
GESTÃO DE DADOS
DE INVESTIGAÇÃO

Pedro Principe

pedro.principe@usdb.uminho.pt