

# Avaliação de infraestrutura para gestão de dados de pesquisa

**LUANA FARIAS SALES, DSc**

Comissão Nacional de Energia Nuclear/

Instituto de Engenharia Nuclear- IEN

PPGCI IBICT-UFRJ

UNIRIO

**LUÍS FERNANDO SAYÃO, DSc**

Comissão Nacional de Energia Nuclear

Centro de Informação Nuclear

PPGB – UNIRIO

**Grupo de Pesquisa: Gestão do**

**Conhecimento Nuclear**



# AGENDA

- ✓ **Considerações Iniciais**
- ✓ **Por que preservar Dados de Pesquisa?**
- ✓ **Como Preservar?**
  - ✓ **Repositório Digital de Dado de Pesquisa**
  - ✓ **Gestão de Dado de Pesquisa**
- ✓ **Avaliação de Infraestrutura de Dados de Pesquisa**
- ✓ **À Guisa de Conclusão**



# DILÚVIO DE DADOS DE PESQUISA

“ DADOS DE PESQUISA SÃO GERADOS POR DIFERENTES COMUNIDADES CIENTÍFICAS E POR MEIO DE DIFERENTES PROCESSOS, PARA DIFERENTES PROPÓSITOS

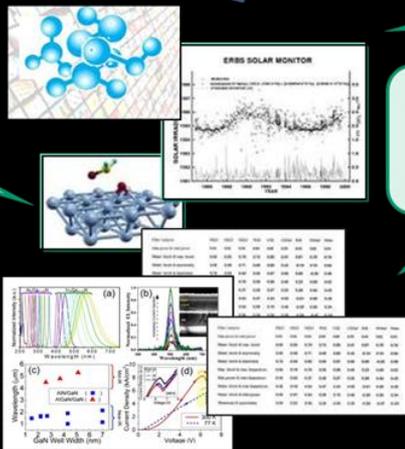
ENTREVISTAS

ESPECIFICAÇÃO DE INSTRUMENTOS OU DE OUTROS *HARDWARE*

CERTIFICADOS DIGITAIS PARA INSTRUMENTOS CIENTÍFICOS

COMENTÁRIOS E ANOTAÇÕES de agentes que tenham consultado os objetos digitais

SOFTWARE código fonte ou implementados como serviços web



FORMULAS MATEMÁTICAS expressas em MathXML ou algoritmos

COLEÇÃO DE DADOS resultados de experimentos, medidas, resultados de levantamentos

VISUALIZAÇÃO DE DADOS gráficos, diagramas, tabelas ou modelos em 3D

ESTRUTURAS QUÍMICAS LÉGIVEL POR MÁQUINA

DOCUMENTOS TEXTUAIS que fazem parte de corpus criados para propósitos de pesquisa

MULTIMÍDIAS imagens, vídeos e gravações em áudio

POLÍTICAS DE AMPLO ESPECTRO

OPÇÕES GERENCIAIS E TECNOLÓGICAS PARA O ARQUIVAMENTO CURADORIA DIGITAL

# ORIGENS DOS DADOS

**DADOS OBSERVACIONAIS** são obtidos de observações diretas, tais como erupção de um vulcão numa data específica, a atitude dos eleitores ou fotografia de uma supernova – que constituem enfim registros históricos que não podem ser coletados uma segunda vez e, portanto, devem ser arquivados para sempre

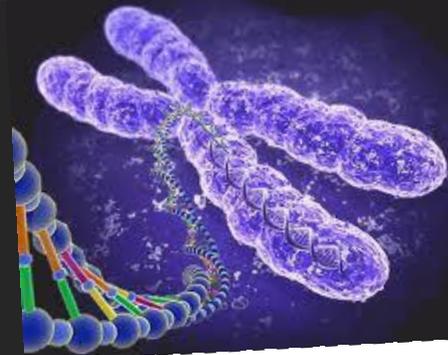


**CRÍTICOS**



**DADOS EXPERIMENTAIS** são provenientes de situações controladas em bancadas de laboratórios. Em tese, dados experimentais provenientes de experimentos que podem ser precisamente reproduzidos e não precisam ser armazenados indefinidamente; entretanto, nem sempre é possível reproduzir precisamente todas as condições experimentais.

**DADOS COMPUTACIONAIS** – resultados da execução de modelos computacionais ou de simulações; devem ser submetidos a uma abordagem distinta que pressupõe o arquivamento de um grande número de informações, expressos por um conjunto robusto de metadados, que incluem descrição de hardware, software e dados de entrada





## **DADOS BRUTOS ou DADOS PRIMÁRIOS**

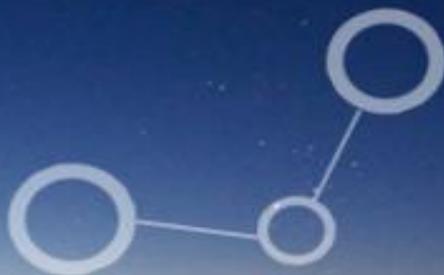
Dados provenientes  
diretamente do  
instrumento científico

.PROCESSAMENTO  
. CALIBRAÇÃO  
.VALIDAÇÃO  
.COMBINAÇÃO COM OUTROS  
DADOS

## **DADOS DERIVADOS**

## **DADOS REFERENCIAIS OU CANÔNICOS**

Coleções de dados consolidados, revisados e geralmente passados por processos de curadoria que estão arquivados em centros de dados. Por exemplo: banco de dados de sequência genética, estruturas química, dados espaciais.



**DINÂMICOS**

**ESTÁVEIS**

**EXPERIMENTAIS**

**OBSERVACIONAIS**

**ÊFEMEROS**

**TERCIÁRIOS**

**SUJOS**

**COMPUTACIONAIS**

**PRIMÁRIOS**

**DERIVADOS**

**LIMPOS**



**TIPOS**

**TEXTO**

**SIMULAÇÃO**

**SOFTWARE**

**VIDEO**

**ENTREVISTA**

**FÓRMULA**

**GRÁFICOS**



**TECNOLOGIAS**

**ASTRONOMIA**

**FÍSICA**

**MEDICINA**

**ECOLOGIA**

**CIÊNCIAS SOCIAIS**

**HISTÓRIA**



**DISCIPLINAS**

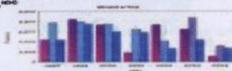
Alguns aspectos da atividade de curadoria variam amplamente de acordo com o **TIPO DE DADOS**, **TECNOLOGIAS SUBJACENTES AOS DADOS**, e, sobretudo, com o **DOMÍNIO DISCIPLINAR ESPECÍFICO**.





# REUSO DOS DADOS DE PESQUISA

demonstrate their results and helps their peers to verify these results. It also makes other researchers aware of the availability of these resources, which may lead to their reuse, saving other researchers the work of e.g. recollecting research data. They also enable creating indirect links between different publications that are possibly related. The Internet provides an infrastructure to publish text with visualizations, animations, research data, etc. Woutersen-Windhouwer and Brandama (2008) indicated several initiatives for publishing enhanced publications on the web, but showed that these initiatives are not easily applicable: they don't fit into existing repository systems, there is little scientific awarding for the additional efforts required for this type of publication and archives do not know how to ingest this material. More generic solutions are needed to overcome these issues.



Data Archiving and Networked Services (DANS) is an institute of both the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organization for Scientific Research (NWO). DANS is responsible for archiving research data from the humanities and social sciences, keeping these data accessible and

	C	D	E	F	G	H
06/23/05	21.02.05	Lot: 0950052.00	Loc: IAMR 0000	Cy		
01	0950052.00	Loc: IAMR 0000	Operator: 365932			
2	LE_THK_TMAX	THK_TE_THK_T	CHORD_T	LE_THK_TMAX	1	
3	0.0091	0.0091	0.0091	0.0156	0.0091	0.01
4	-0.0001	-0.0001	-0.0001	-0.0156	-0.0001	-0.01
5	-0.0001	-0.0001	-0.0001	-0.0161	-0.0001	-0.01
6	0.0045	-0.0024	0.0036	0.0079	0.0059	-0.01
7	0.0056	0.0096	0.0096	0.0161	0.0096	0.01
8	6/27/2005	0.07.13	0.0026	-0.003	0.0036	0.0047
9	6/27/2005	0.07.22	0.0014	-0.0034	0.0033	0.0056
10	6/27/2005	0.07.32	0.0032	-0.0038	0.0016	-0.0019
11	6/27/2005	0.07.39	0.0027	-0.0029	0.0039	0.0019
12	6/27/2005	0.07.59	0.0028	-0.0038	0.0022	0.0012
13	6/27/2005	0.08.08	0.0012	-0.0045	0.0022	0.0025
14	6/27/2005	0.08.15	0.004	-0.0019	0.0036	0.0103
15	6/27/2005	0.08.24	0.0031	-0.0029	0.0034	0.0038
16	6/27/2005	0.08.30	0.0031	-0.0028	0.0036	0.0092

GESTÃO & CURADORIA

ANALISADOS  
EM NOVOS E  
DIFERENTES  
CONTEXTOS

Os pesquisadores começaram a creditar **toda a confiança** nos conteúdos digitais **criados por outros pesquisadores** para dar prosseguimento aos seus empreendimentos

# REUSO DE DADOS DE PESQUISA

## EM OUTROS CONTEXTOS

### LIMITES TEMPORAIS

diários de bordos de navios do século XVII digitalizados e depois analisadas por climatologista do século XX

### LIMITES SETORIAIS

epidemiologistas examinam dados comerciais sobre consumo em busca de remédios para a gripe

### LIMITES DISCIPLINARES

pesquisadores em bioinformática combinam coleções de dados originados no domínio da biologia, genética e engenharia

**A probabilidade de uma coleção de dado ser reusada no futuro por outras audiências, estabelece o critério mais simples de valor para a coleção; embora não seja algo simples, a partir daí pode-se estimar se vale pena arquivá-la por longo prazo**



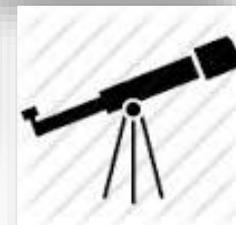
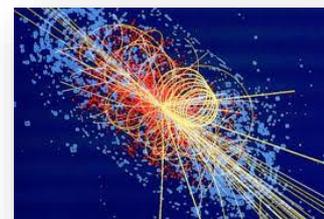
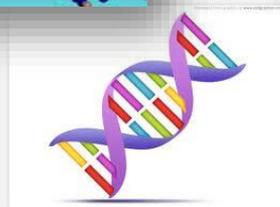
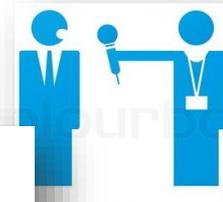
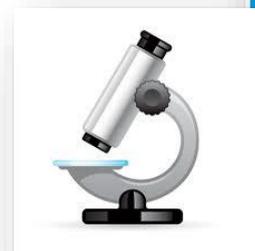
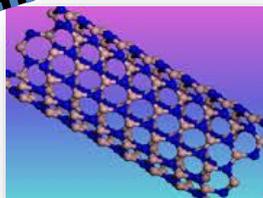
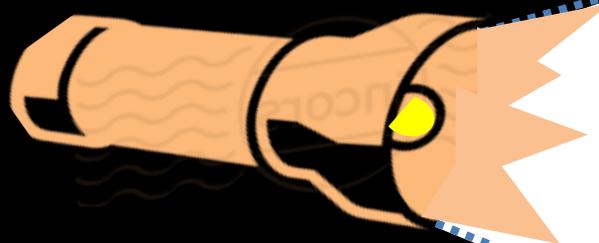
Há uma parcela dos produtos de pesquisa que necessita de infraestruturas

**INFORMACIONAIS**

**TECNOLÓGICAS**

**POLÍTICAS**

**GERENCIAIS**



Para se tornarem visíveis para as comunidades acadêmicas, Instituições de pesquisa, agências de fomento e para o cidadão comum.

**DADOS DE PESQUISA  
SE TORNAM PARTE DA  
INFRAESTRUTURA  
MUNDIAL DE PESQUISA**

**AGÊNCIAS FINANCIADORAS DE  
PESQUISA**

**PLANOS DE COMPARTILHAMENTO DE DADOS  
POLÍTICAS MANDATÓRIAS**

Isso garante que os **pesquisadores se comprometem a cuidar dos dados** durante e após a pesquisa no sentido de otimizar o compartilhamento de dados.

**PERIÓDICOS CIENTÍFICOS**

Os periódicos exigem cada vez mais que os dados que sustentam a pesquisa publicada depositado dentro em uma **base de dados ou repositório** acessível .

**INSTITUIÇÕES DE PESQUISA**

**Financiam/desenvolvem infraestruturas** para gestão e serviços de dados para facilitar o compartilhamento dentro de domínios específicos.

**PESQUISADORES**

Iniciativas como o DataCite - que atribui identificador persistente (DOI) aos dados de pesquisa - ajudam o cientista a tornar seus dados citáveis, rastreáveis e acessíveis de modo que os dados de pesquisa, bem como as publicações baseadas nesses dados, façam parte da produção científica desse pesquisador.



## CALL FOR RESEARCH PROPOSALS - ESCIENCE 2015

### Characteristics of the research proposals

**Data management plan:** A major characteristic of eScience projects is its dependency on data management practices, and the **need of making results public, to allow reuse and collaboration with other groups**. Therefore, all projects should provide indication of how they intend to manage the data produced during the project (where the term “data” is taken on the large, and includes files, algorithms, software, samples, models, curriculum material and others).

# nature.com

The world's best science and medicine on your desktop

## Availability of data, material and methods

An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. **Therefore, a condition of publication in a Nature journal is that authors are required to make materials, data and associated protocols promptly available to readers without undue qualifications.** Any restrictions on the availability of materials or information must be disclosed to the editors at the time of submission. Any restrictions must also be disclosed in the submitted manuscript, including details of how readers can obtain materials and information. If materials are to be distributed by a for-profit company, this must be stated in the paper.





## DEPOSITAR & COMPARTILHAR

infraestruturas que assegurem o máximo de **confiabilidade, estabilidade e acessibilidade** e que facilitem o trabalho de **arquivamento, compartilhamento** e **reconhecimento de autoria** para os seus dados

## O QUE PRECISAM OS PESQUISADORES?

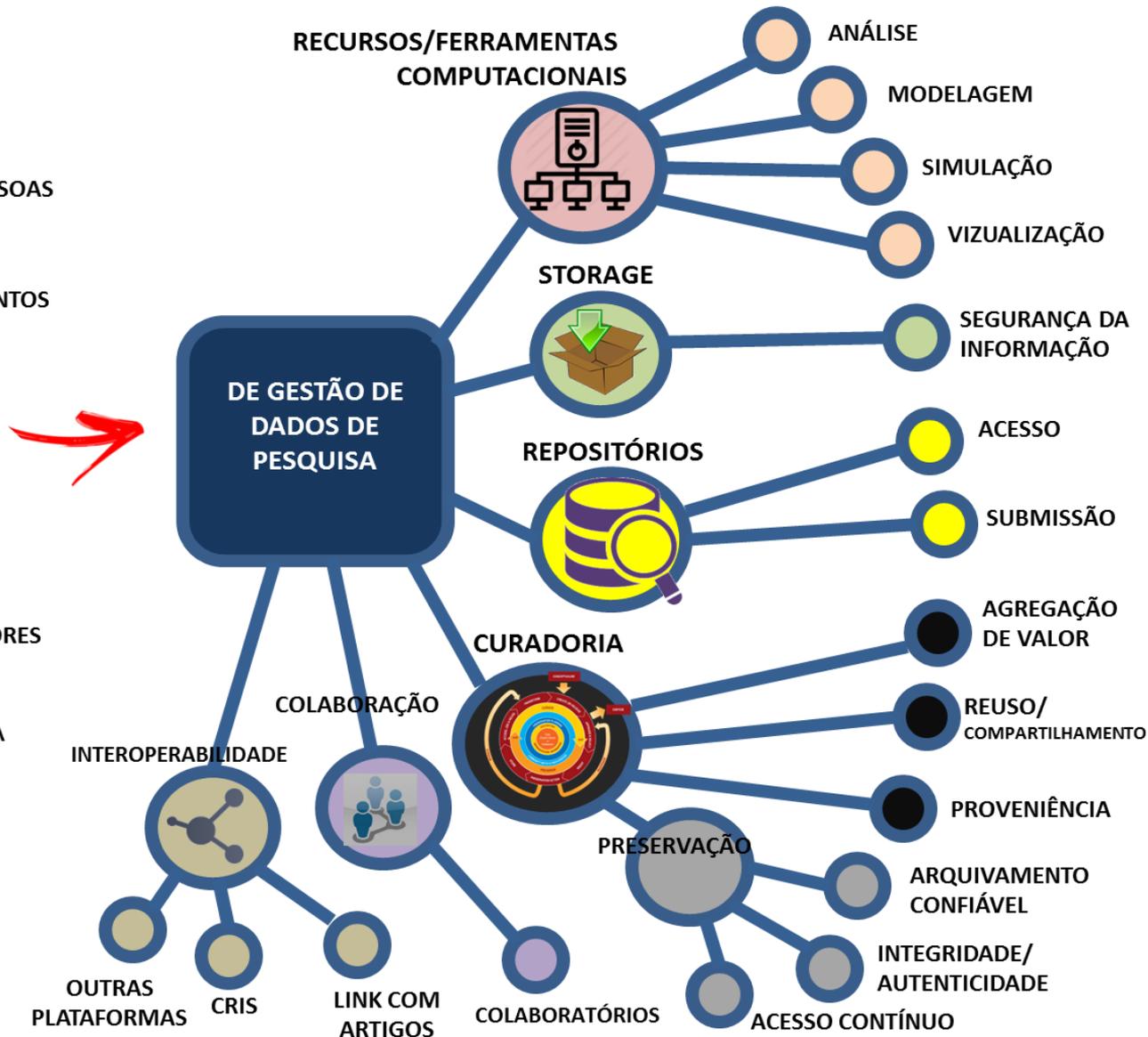
[ e os outros interessados ]

## DESCOBRIR E ACESSAR

precisam **encontrar coleções de dados** de pesquisa, saber como acessá-las e sob que condições podem reutilizar esses dados e assim dar prosseguimento às suas pesquisas **confiando na autenticidade e proveniência dos dados coletados ou gerados por outros pesquisadores.**

# CIBERINFRAESTRUTURA DE DADOS DE PESQUISA

## FONTES DE DADOS

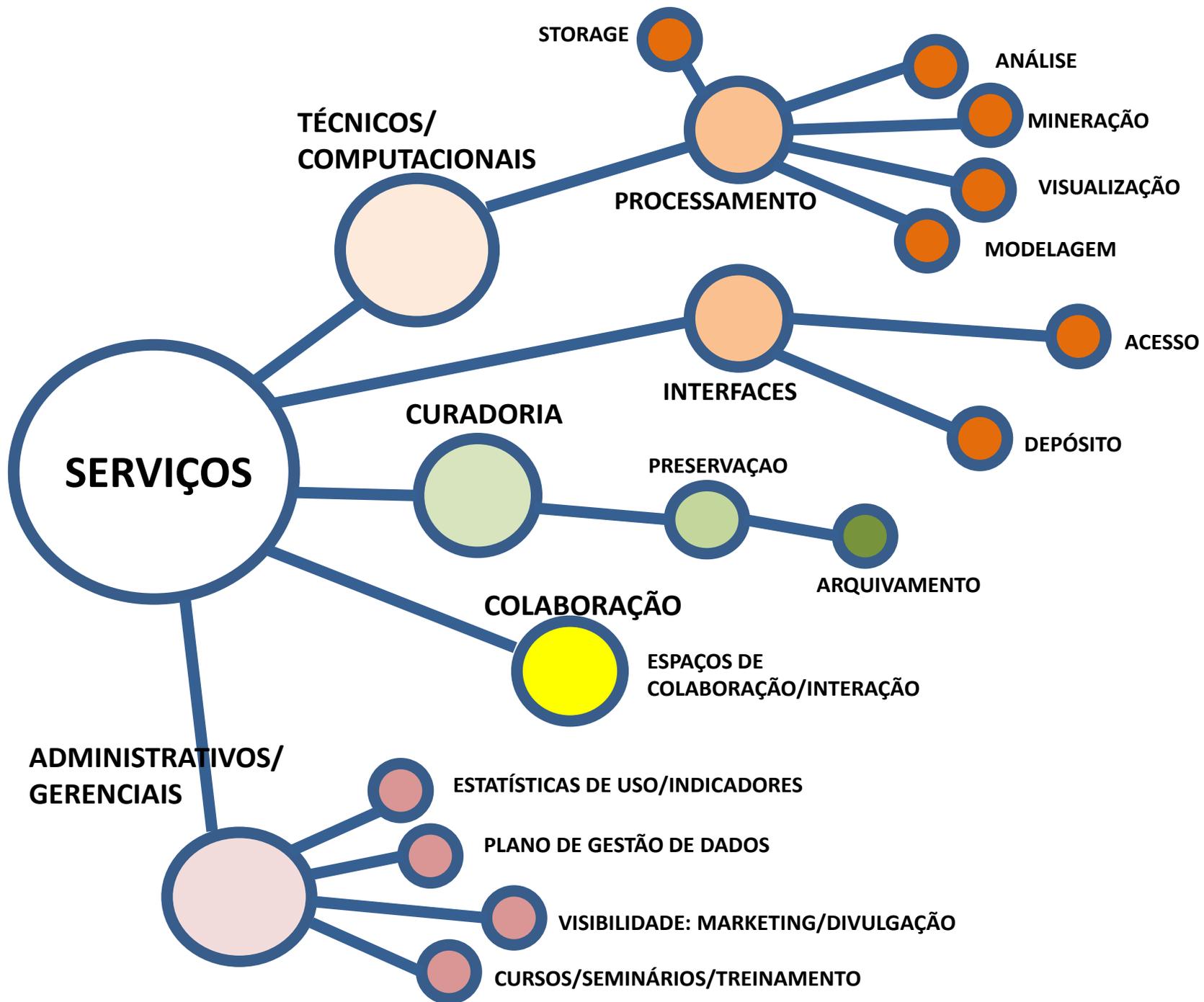


## POLÍTICA DE DADOS DE PESQUISA

# REPOSITÓRIO DIGITAL DE DADOS DE PESQUISA

A primeira exigência para a gestão e curadoria é o dado estar disponível em um repositório





# **DADO DE PESQUISA MANEIRO!**



**ARQUIVADO  
PRESERVADO**

**LOCALIZADO  
RECUPERADO  
ACESSADO**

**INTERPRETADO  
CONTEXTUALIZADO  
AVALIADO  
PROVENIÊNCIA**

**REUSADO**

**COMPARTILHADO  
ON-LINE**

**ANOTADO  
ATIVA COLABORAÇÃO**

**INTEROPERÁVEL**

**LINKADO COM  
PUBLICAÇÃO**

**LICENÇA APROPRIADA**

**CONSIDERA PRIVACIDADE/ÉTICA**

**IDENTIFICADO  
CITADO  
VISÍVEL**

# TRANSMITIR CONHECIMENTO!

1

## ACESSÍVEL

Localizado e acessado

2

## INTELIGÍVEL

Deve se apresentar de forma inteligível para aqueles que desejam entendê-lo ou analisá-lo. Deve ser diferente para diferentes audiências, mesmo para o cidadão comum.

3

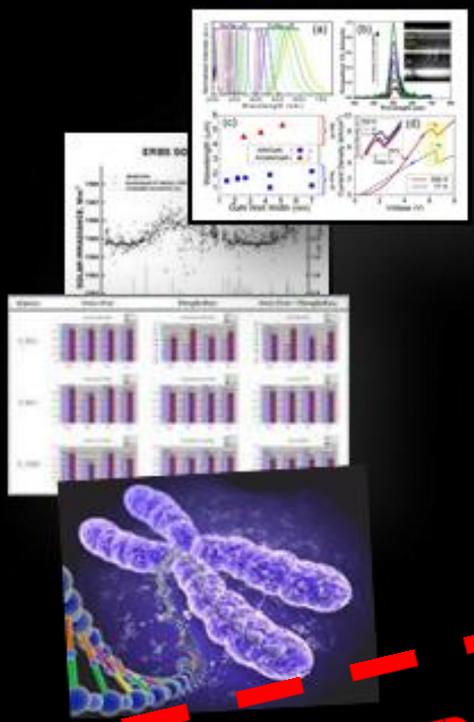
## AVALIÁVEL

Natureza, evidências, confiabilidade, fonte competente, financiador, objetivos

4

## UTILIZÁVEL

Deve ser capaz de ser reusado em diferentes contextos e, minimamente, por outros cientistas.



INTEROPERÁVEL!

NOS ESTAMOS NOS AFOGANDO EM **DADOS**, MAS SEDENTOS DE **INFORMAÇÃO**

**GESTÃO/CURADORIA**



# GESTÃO DE DADOS DE PESQUISA

Em comparação com os repositórios de *e-prints* as funções, as descrições, os padrões e os controles são mais numerosos e complexos. Essa complexidade, no entanto, varia de acordo com os ambientes disciplinares considerados e com a política adotada pela instituição.

AS BIBLIOTECAS DE PESQUISA TEM QUE CAPTURAR DADOS EM DIFERENTES ESTÁGIOS DA GERAÇÃO E PROCESSAMENTO DOS DADOS DE PESQUISA.

## PÓS-PUBLICAÇÃO → PRÉ-PUBLICAÇÃO



WORKFLOW; INSTRUMENTOS;  
MODELOS; FERRAMENTAS;  
CÓDIGOS; CADERNO DE  
PESQUISA...

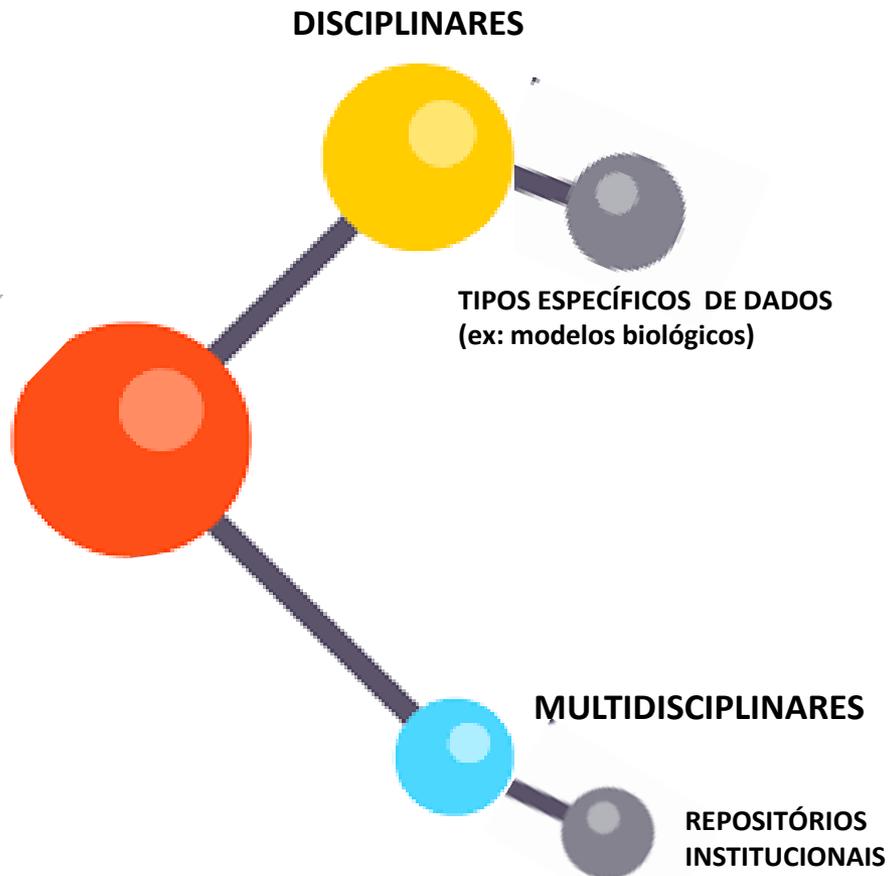
O PLANEJAMENTO DA GESTÃO DE DADOS SE TORNA PARTE DO PROCESSO DE INVESTIGAÇÃO CIENTÍFICA

**O QUE  
AVALIAR?**



A POLÍTICA DE UM REPOSITÓRIO delinea os compromissos que a instituição se obriga em relação aos seus principais *stakeholders* – pesquisadores, curadores, consumidores, financiadores, coletores de metadados, etc. - e com relação ao ciclo de vida das coleções de dados que estão sendo gerenciados.

# TIPOS DE PLATAFORMAS DE GESTÃO DE DADOS

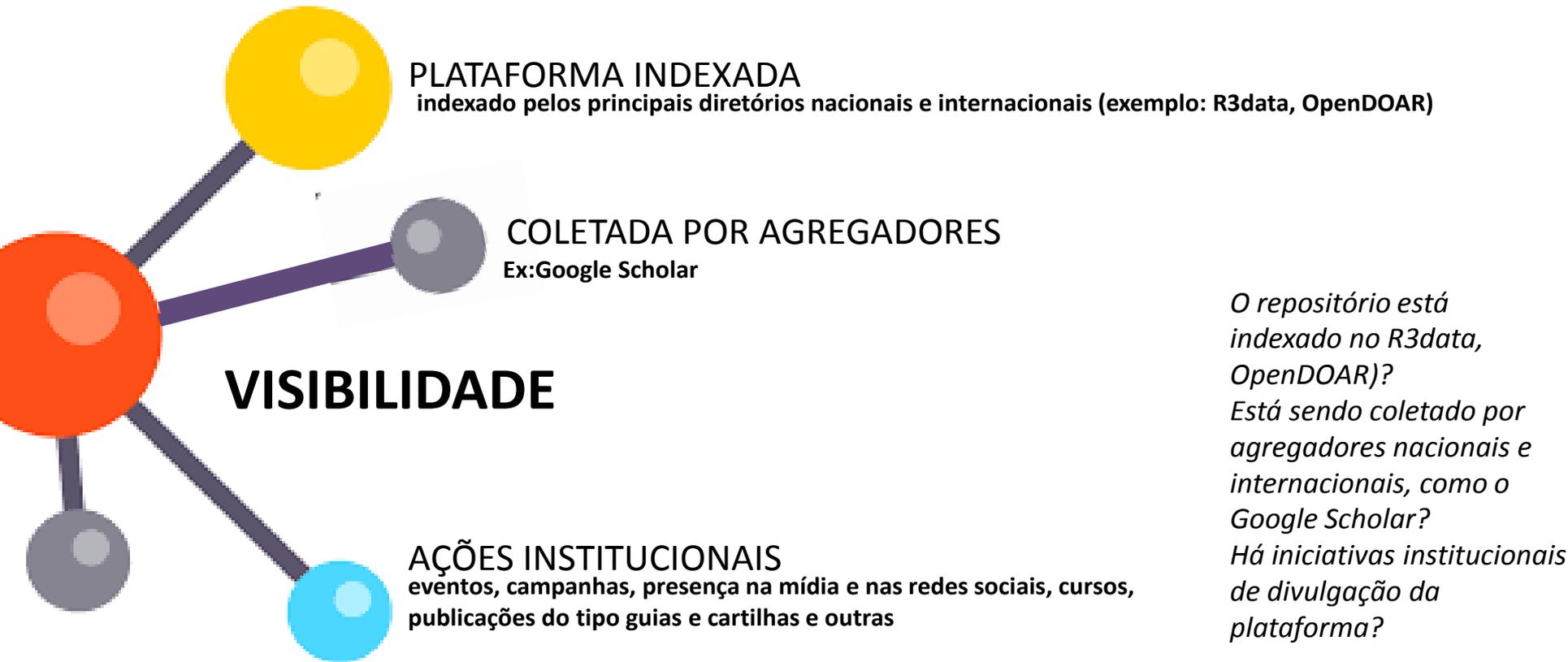


As **PLATAFORMAS DISCIPLINARES** se voltam para domínios específicos ou para tipos particulares de dados. Em geral possuem modelos de dados adequados à representação das coleções de dados e oferecem uma **CARTEIRA DE SERVIÇOS** mais orientadas, como curadoria e visualização.

Essas plataformas estão abertas para publicar qualquer tipo de dados, e são especialmente desenvolvida para dar apoio a publicação de datasets produzidas no âmbito da ciência chamada de **“CAUDA LONGA”** – domínios científicos nos quais um grande número de relativamente pequenos laboratórios ou de pesquisadores individuais produzem a maioria resultados científicos



A falta de **VISIBILIDADE** de informações sobre coleções de dados de pesquisa coloca um grave obstáculo para o acesso e reuso desses recursos. Isto implica em dizer que as informações sobre as coleções de dados, sobre as organizações produtoras, a documentação sobre os dados e as especificações sobre as condições de uso dessas coleções devem estar disponíveis em escala internacional de forma transparente e preferencialmente via internet (OCDE, 2007).



*O repositório está indexado no R3data, OpenDOAR)?  
Está sendo coletado por agregadores nacionais e internacionais, como o Google Scholar?  
Há iniciativas institucionais de divulgação da plataforma?*

RECONHECIMENTO  
DA COMUNIDADE  
ACADÊMICA



# REFERÊNCIA

A capacidade das coleções de dados e suas versões hospedadas nos repositórios de serem **IDENTIFICADAS** permanentemente torna-se essencial para o **acesso, preservação e citação**; é um fator importante também nos processos de **interoperabilidade** e de **linking** com outros recursos via, por exemplo, *linked data*.

## IDENTIFICADORES PERSISTENTES

DOI

URN

HANDLES

Específicos

## CONTROLE DE VERSÕES

UFG – UNIVERSAL FINGERPRINT

TIMESTAMPING

## CITAÇÃO PADRONIZADA

FERRAMENTAS DE APOIO À CITAÇÃO

EXPORTAÇÃO EM FORMATOS DIVERSOS/COMPARTILHAMENTO

O controle de versões é um processo importante para o fundamento da reprodutibilidade da pesquisa, para a integridade da referência às coleções de dados e para proveniência dos seus conteúdos. Isto por que as coleções de dados podem evoluir no tempo por vários motivos



04

27

56

01

16

44

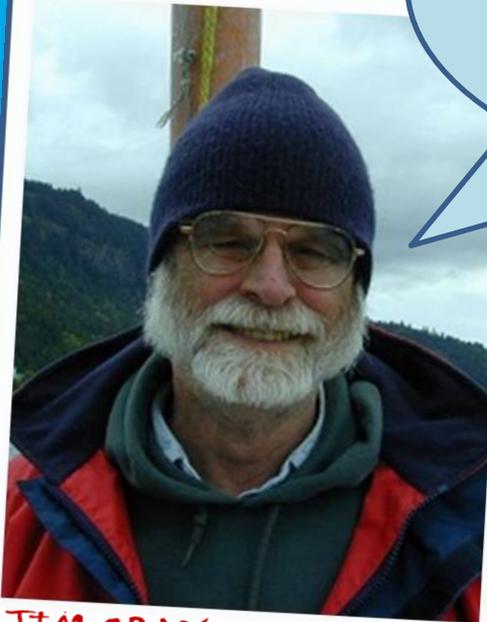
02

01

17

# DADO DE PESQUISA NÃO FALA POR SI PRÓPRIO

Dados de pesquisa são  
incompreensíveis e portanto  
inúteis a menos que haja uma  
descrição detalhada e clara de  
como e quando eles foram obtidos e  
de como os dados derivados foram  
produzidos !!!



**JIM GRAY**  
Cientista da Computação  
Desaparecido em 2007



**POR QUE?**

**QUEM?**

**O QUE?**

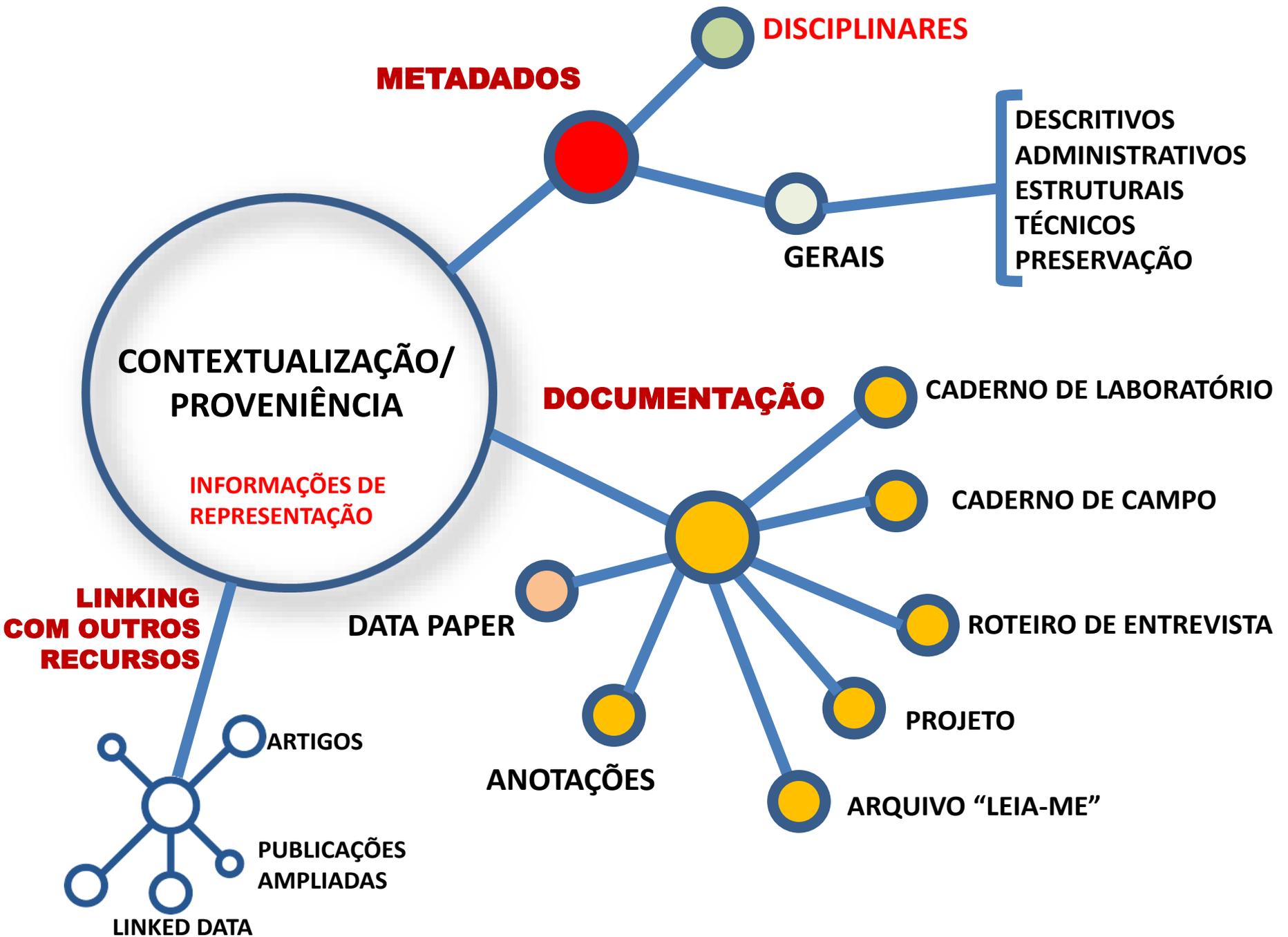
**COMO?**

**QUANDO?**

**ONDE?**



**SIGNIFICADO**  
**ESTRUTURA**  
**IDENTIFICAÇÃO**  
**CONTEXTO**  
**PROVENIÊNCIA**

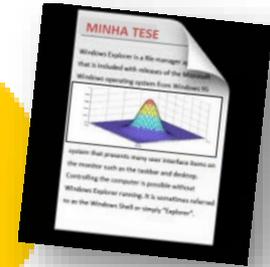




# INTEROPERABILIDADE POSSÍVEL

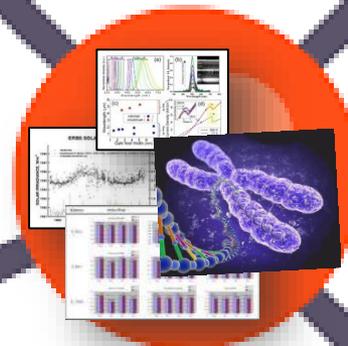


SISTEMAS DE PERIÓDICOS

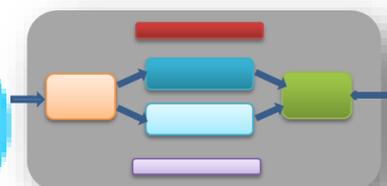


REPOSITÓRIOS INSTITUCIONAIS

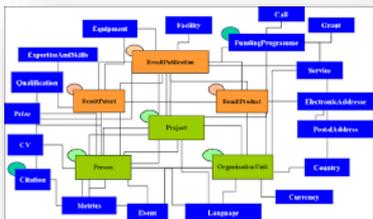
REPOSITÓRIOS DE DADOS



REPOSITÓRIOS CONFIÁVEIS



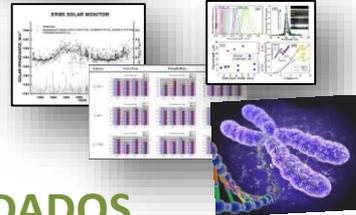
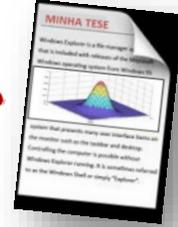
CLOCKSS



SISTEMAS CRIS  
GESTÃO DE PESQUISA

# INTEGRAÇÃO OJS x DATAVERSE

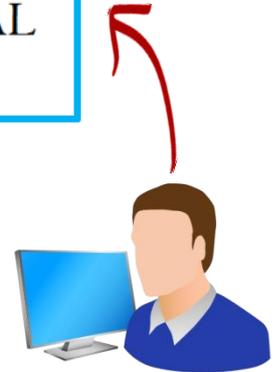
ARTIGO



DADOS



CITAÇÃO  
INCLUIDA  
NO ARTIGO



REVISOR

DADOS  
METADADOS  
DOCUMENTAÇÃO

API



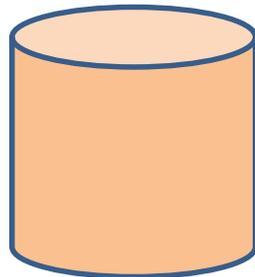
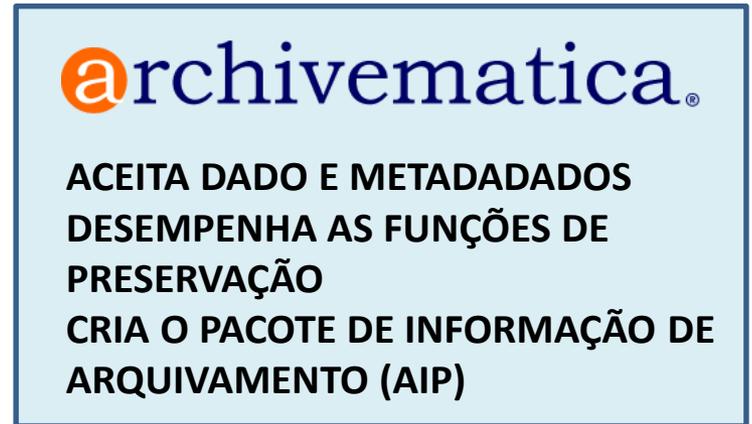
LINK PARA O ARTIGO

DOI

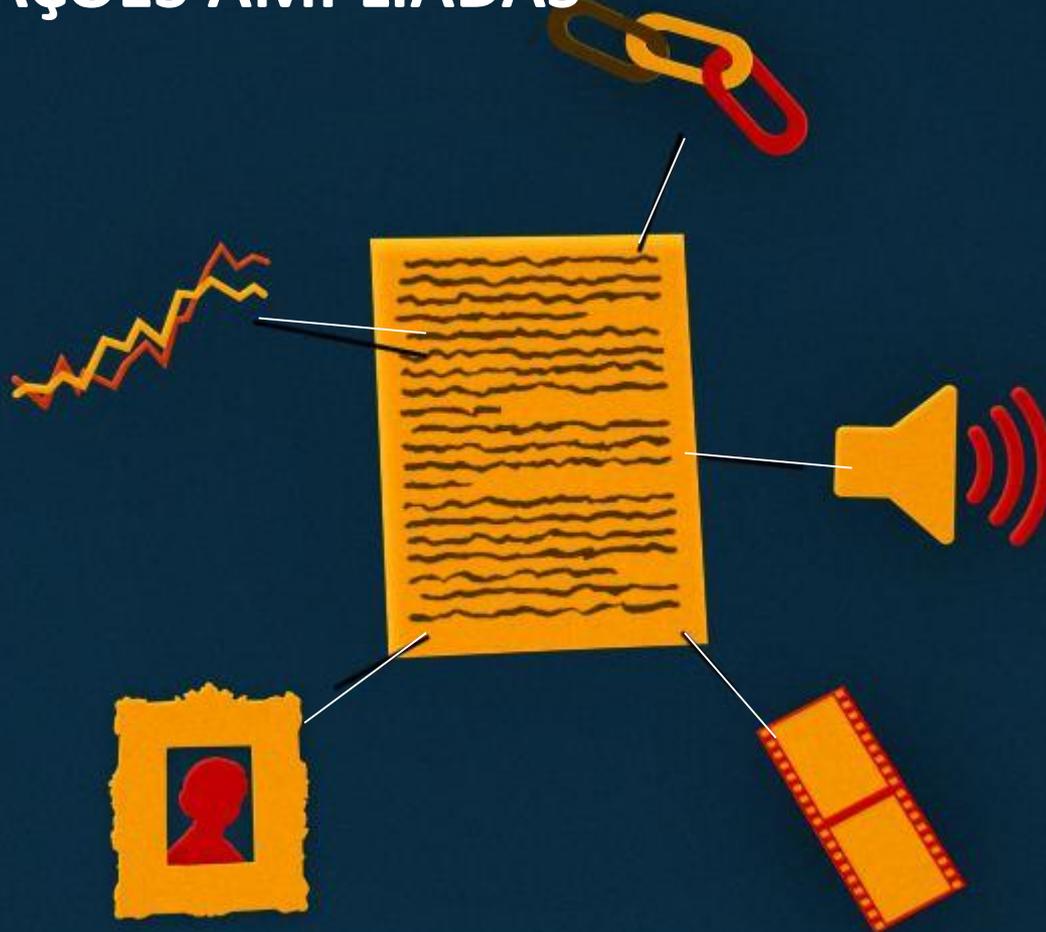
# INTEGRAÇÃO DATAVERSE X ARCHIVEMATICA



API ou SWORD



# PUBLICAÇÕES AMPLIADAS



**SISTEMAS CORPORATIVOS**

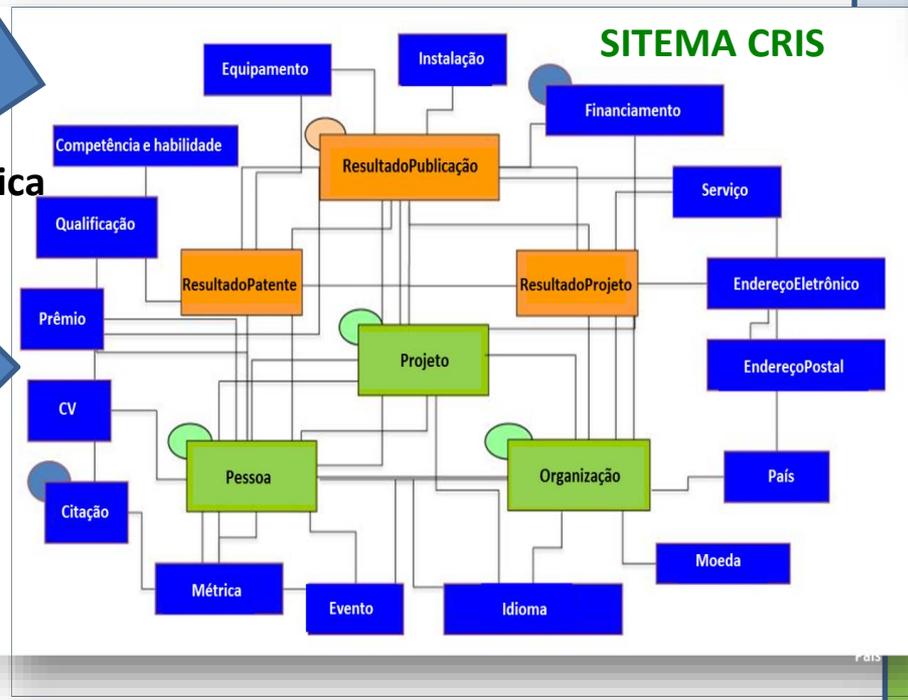
**RH**  
**Sistema de gestão acadêmica**

**OUTROS SISTEMAS CRIS**

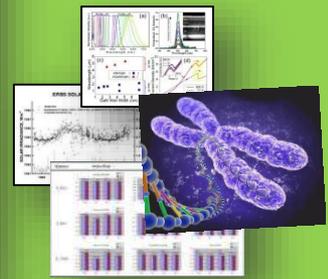
**CERIF**

**OUTROS SISTEMAS**

**LATTES**  
**WEB OF SCIENCE**  
**arXiv**



**REPOSITÓRIOS INSTITUCIONAIS**



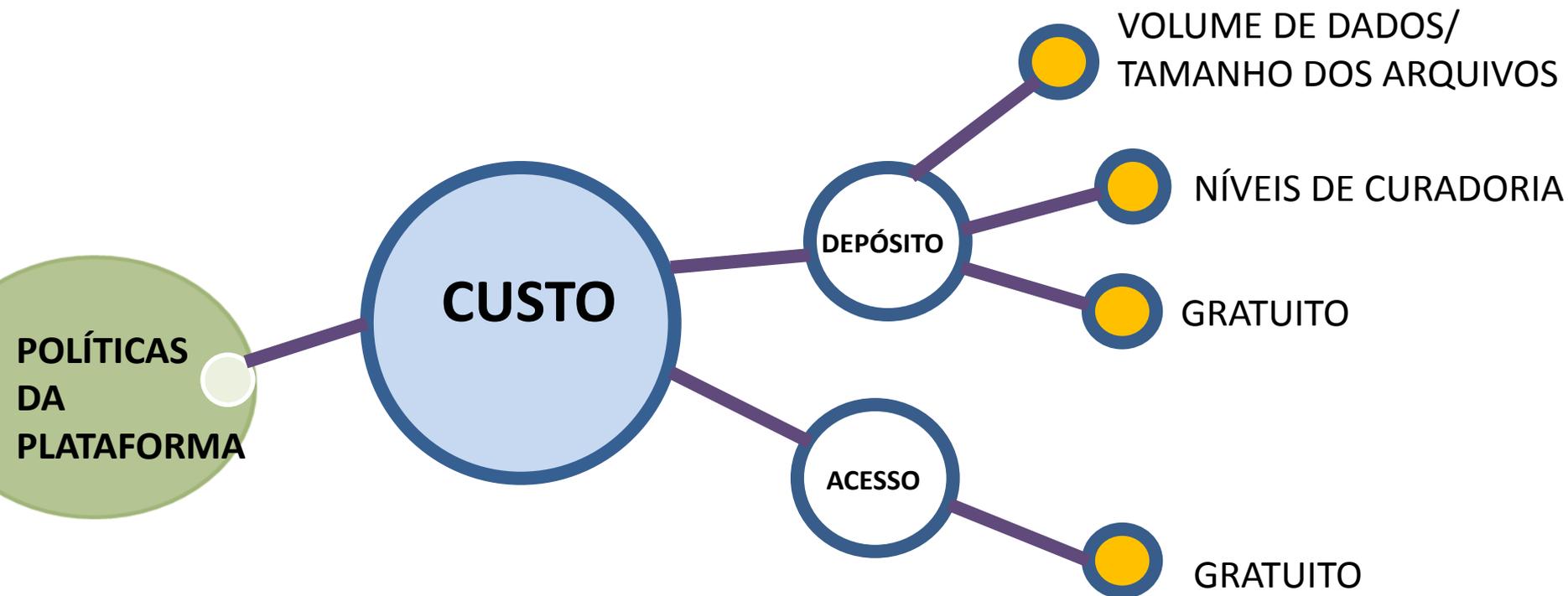
**REPOSITÓRIOS DE DADOS**

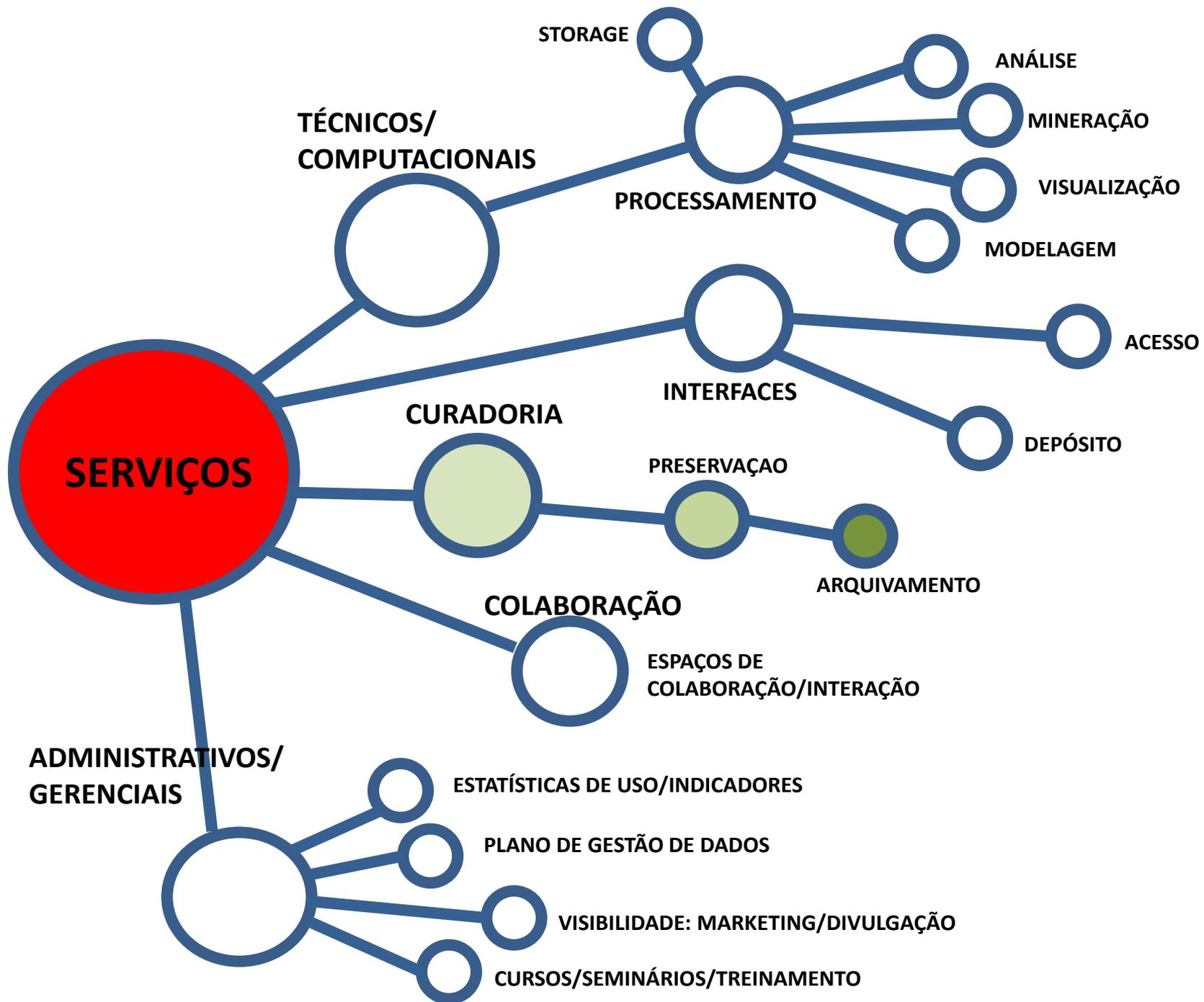
# CERTIFICAÇÃO & AUDITORIA

A IMPORTÂNCIA DOS PROCESSOS DE AVALIAÇÃO RESIDE NO FATO DELES PROMOVEREM A CONFIANÇA NA USABILIDADE, SUSTENTABILIDADE E PERSISTÊNCIA POR LONGO PRAZO DOS DADOS DISPONÍVEIS PARA COMPARTILHAMENTO.



A operação serviços de gestão de pesquisa pressupõe um custo considerável - tanto monetário quanto de custos de outra natureza - para as instituições que abrigam estas plataformas. **Esse custo está entre os principais fatores que impedem que a publicação de dados de pesquisa seja uma norma corrente na ciência.**





# CURADORIA DIGITAL

# CURADORIA



TRADICIONALMENTE CURADORIA NÃO ESTÁ PREOCUPADA APENAS COM A PRESERVAÇÃO POR LONGO PRAZO DE LIVROS RAROS, PINTURAS E OUTROS ARTEFATOS. SEU FOCO PRINCIPAL É MANTER A INTEGRIDADE E POSSIBILITAR E PROMOVER SUA DISPONIBILIDADE PARA AUDIÊNCIAS APROPRIADAS.

**ISTO TAMBÉM É VERDADEIRO PARA DADOS DE PESQUISA. PARA QUE ELES PERMANEÇAM ÚTEIS PODE SER NECESSÁRIO MANUTENÇÃO E MELHORIAS. ISTO SOMADO A PROMOÇÃO DOS DADOS PARA OS CONSUMIDORES POTENCIAIS SÃO OS DOIS PAPÉIS PRINCIPAIS DA CURADORIA DE DADOS.**

**À GUIZA DE CONCLUSÃO**

- Precisa de Tecnologia? Sim!! Precisa muito mais de gestão
- É um campo riquíssimo de pesquisa, pois a agenda se renova, valorizando as bases teóricas da Biblioteconomia, Arquivologia e Ciência da Informação
- É uma área rica por sua interdisciplinaridade
- Reaproxima as bibliotecas de pesquisa dos laboratórios e conseqüentemente do mundo da pesquisa
- É uma área complexa? Sim, mas pode se começar pequeno até chegar ao todo mais complexo (sempre atento aos padrões)

