CURADORIA DIGITAL & DADOS DE PESQUISA

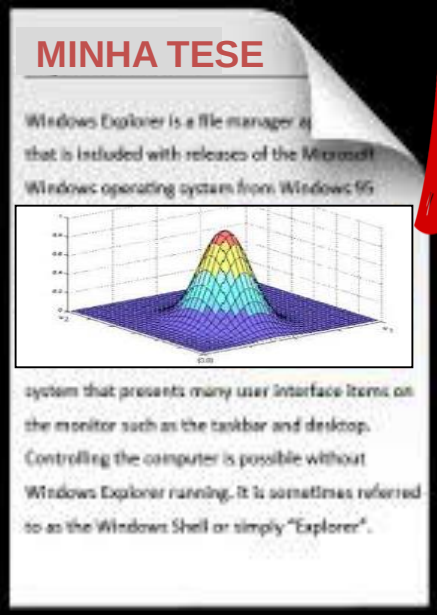
Luis Fernando Sayão
CNEN/CIN

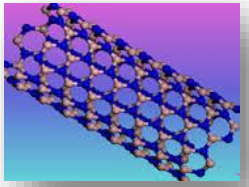
NA CAUDA LONGA DA CIÊNCIA

VISÍVEL

INVISÍVEL

O TEXTO ACADÊMICO APRESENTA APENAS OS DADOS DE PESQUISA DE FORMA CONDENSADA

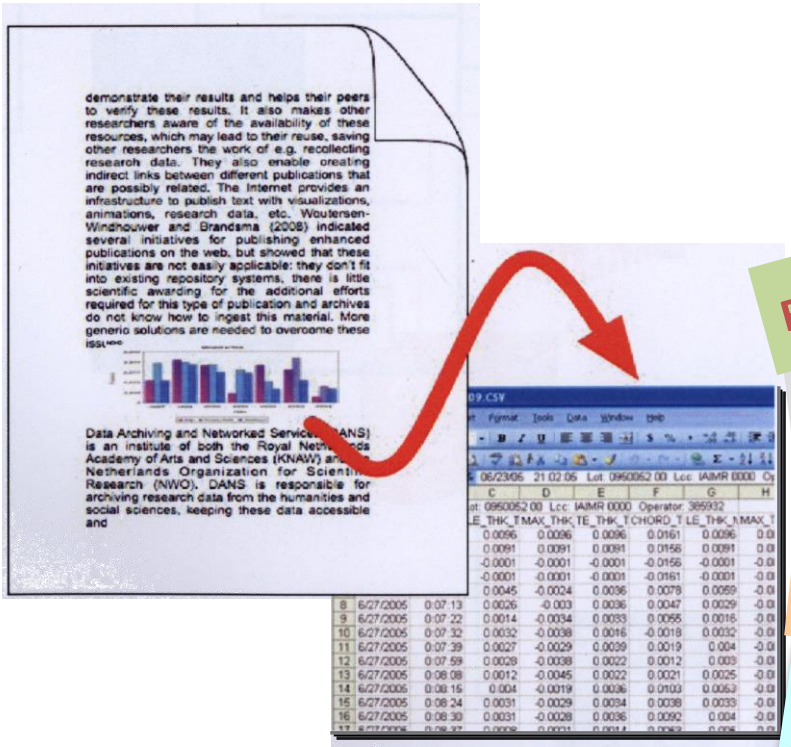




UMA VISÃO DOS DADOS !!!

[revisão por pares]
[validação da pesquisa]

OS DADOS SUBJACENTES ÀS PUBLICAÇÕES APOIAM A:



REPRODUTIBILIDADE

VALIDAÇÃO

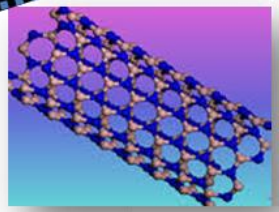
AUTOCORREÇÃO

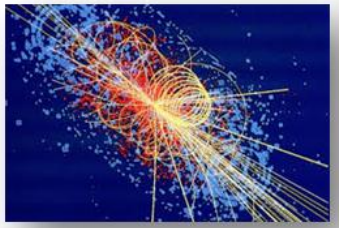
AVALIAÇÃO POR PARES

INTERPRETAÇÃO

BASE PARA NOVAS PESQUISAS

Há uma parcela dos produtos de pesquisa que necessita de infraestruturas
INFORMACIONAIS
TECNOLÓGICAS
POLÍTICAS
GERENCIAIS



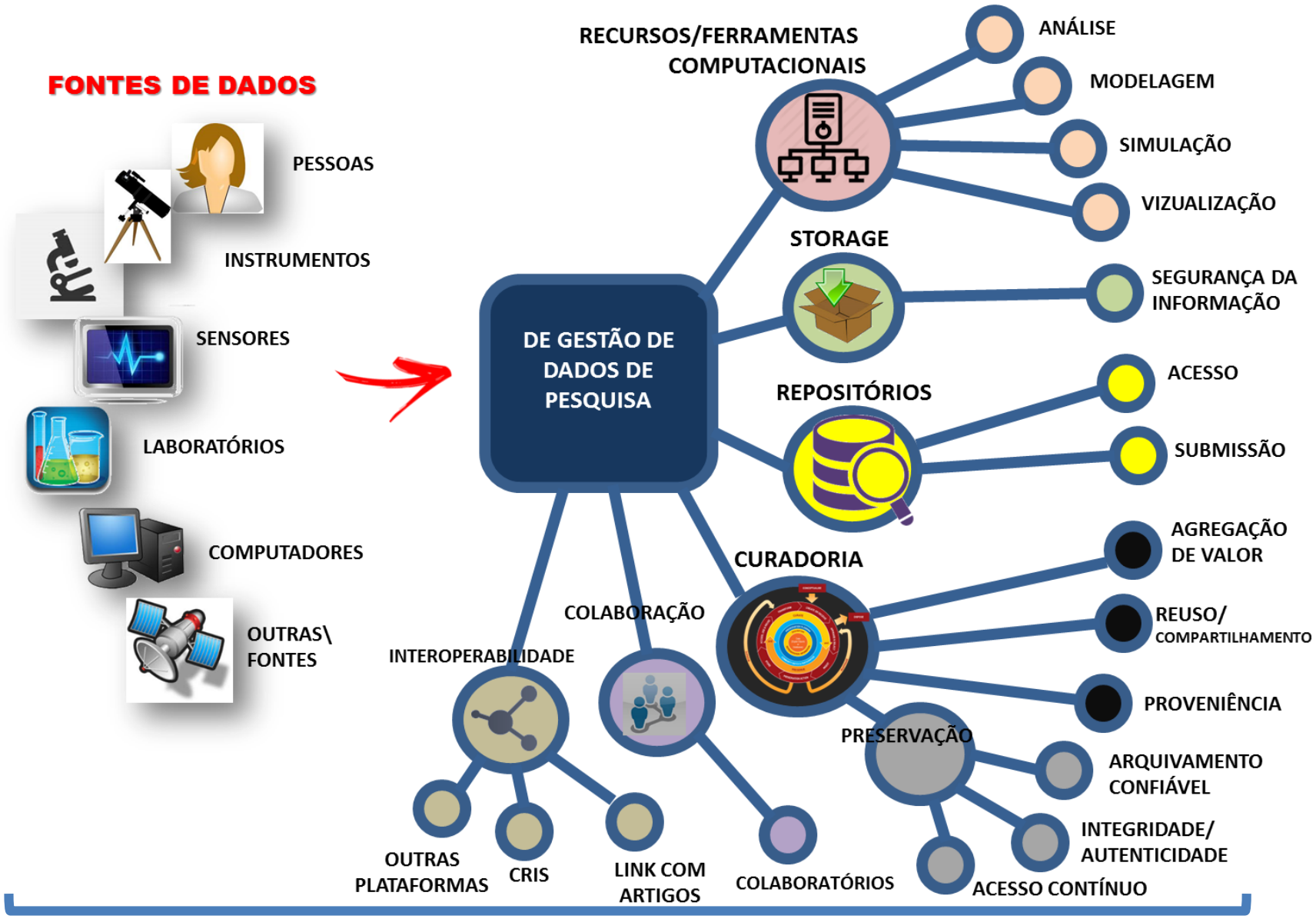


Para se tornarem
visíveis para as comunidades acadêmicas, Instituições de pesquisa, agências de fomento e para o cidadão comum.


CIDADÃO COMUM
Transparência nas atividades de pesquisa
Compreender os benefícios da ciência
abrir a caixa preta da ciência
AGÊNCIA DE FOMENTO
Otimização de recursos públicos
Duplicação de esforços
Validação das pesquisa
CIÊNCIA ABERTA
Expansão do conceito de acesso livre
para dados, metodologia,
códigos, cadernos de laboratório
eSCIENCE
Significado, compartilhamento,
análises, pesquisa distribuída
Pesquisas interdisciplinares
INSTITUIÇÃO DE PESQUISA
Incorporar os dados na memória
Acadêmica. Reuso dos seus dados
Indicador de produtividade
PESQUISADOR
Reconhecimento, Citação,
Recompensa por preparar e
publicar os dados
EDITOR CIENTÍFICO
Expandir seus modelos de
negócio
PROFESSOR
Usar dados no ensino de ciências
DEMANDAM
POLÍTICA NACIONAL
DE DADOS PESQUISA
POLITICAS
INSTITUCIONAIS
GESTÃO DE DADOS
DE PESQUISA
INFRAESTRUTURAS para
ORGANIZAÇAO
DOCUMENTAÇÃO
ARQUIVARMENTO
PRESERVAÇAO
CURADORIA
DESCOBERTA
COMPARTILHAMENTO
SEGURANÇA


Políticas
Financiamento
Capacitação
Pesquisa
Legislação
Boas práticas
Parcerias
INSTITUCIONALIZAÇÃO
Formato dos dados
Metadados
Identificação persistente
Preservação/curadoria
Ontologias/taxonomias
Protocolos de interoperabilidade
Segurança
Interoperabilidade
PADRÕES
TECNOLOGIAS
Redes de computadores
Banco de dados
Ferramentas de software
Sistemas de *storage*
Repositórios confiáveis

CIBERINFRAESTRUTURA DE DADOS DE PESQUISA



POR QUE GERENCIAR?

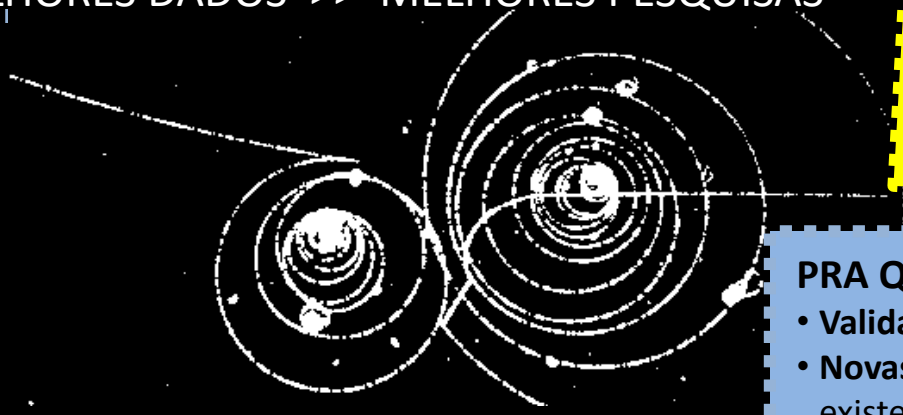
- Preservar a **integridade da pesquisa**
- Permitir que os dados estejam **disponíveis** para ser usados por outros;
- Ajudar o pesquisador a reduzir o risco de **perdas de dados;**
- Assegurar o **acesso contínuo** aos dados de valor.

POR QUE CONECTAR OS DADOS?

- Interligar dados a pessoas, a projetos e a publicações;
- Aumentar a **encontrabilidade** dos dados;
- Conectar dados a **descobertas científicas;**
- Promover um **contexto** rico para os dados.

DESAFIOS !

MELHORES DADOS >> MELHORES PESQUISAS



POR QUE TORNAR OS DADOS ENCONTRÁVEIS?

- Possibilitar que se demonstre a excelência da pesquisa;
- **Permitir que os pesquisadores desenvolvam novas pesquisas sobre os dados existentes, ao invés de recriá-los;**
- Promover a **inovação**;
- Proporcionar a capacidade de resolver grandes problemas de forma **interdisciplinar**.

PRA QUE REUSAR OS DADOS?

- **Validar** as pesquisas;
- **Novas descobertas** baseadas nos dados existentes;
- **Integração de coleções** de dados para novas análises;
- **Reanálise** de pesquisas caras, raras ou que não podem ser repetidas;
- **Redução de duplicação** de esforços

ORIENTADA POR DADOS

BIG SCIENCE

GRANDES INSTRUMENTOS
ALTOS CUSTOS
LONGA DURAÇÃO
MUITOS COLABORADORES
PESQUISA DISTRIBUÍDA

SMALL SCIENCE

PEQUENOS INSTRUMENTOS
BAIXOS CUSTOS
PEQUENA DURAÇÃO
EQUIPES PEQUENAS
PESQUISA LOCAL

ORIENTADA POR HIPÓTESES

A GRANDE CIÊNCIA

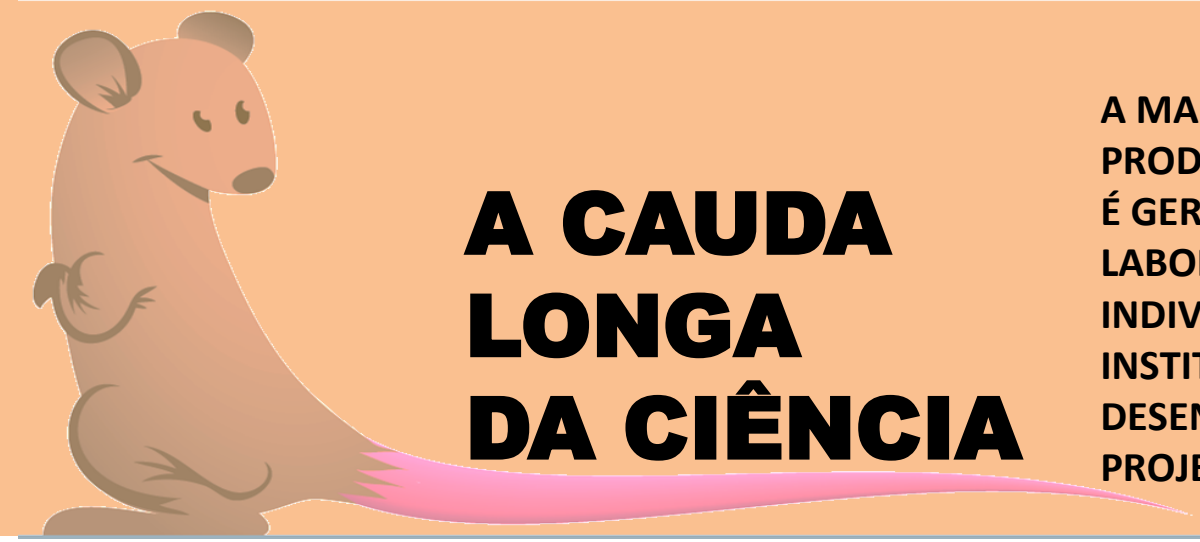
O aspecto de grande escala da ciência moderna – nova, brilhante e todo-poderosa – é tão evidente que foi criado, para descrevê-la, o expressivo termo "**Grande Ciência** [...]

A Grande Ciência é tão recente que muitos de nós não recordamos de suas origens. A grande ciência é tão vasta que começamos a nos preocupar com o tamanho do monstro que criamos. A grande ciência é tão diferente da anterior, que nos lembramos, talvez com nostalgia da "Pequena Ciência" que constituiu no passado nossa maneira de viver (PRICE, 1976, p.2.)

Nos dias de hoje, com maior ímpeto, **os investimentos e as atenções dos órgãos formuladores das políticas científicas se voltam para esses segmentos da ciência contemporânea.** Essa nova configuração colossal que se instala no cenário científico, têm sido comparados as pirâmides do Egito e as grandes catedrais da Europa Medieval (WEINBERG, 1961).

CARACTERÍSTICAS		CABEÇA	CAUDA LONGA
UNIFORMIDADE	DIVERSIDADE	homogêneos	heterogêneos
	GERAÇÃO/ COLETA	instrumentos automatizados	gerados/coletados manualmente
	PROCEDIMENTOS	padronizados	específicos
GESTÃO	CURADORIA	Centralizada/ institucionalizada	Individual
	REPOSITÓRIOS DIGITAIS	Disciplinares ou referenciais	Institucionais ou Multidisciplinares
	PRESERVAÇÃO	Preservados	Não preservados
	ARMAZENAMENTO	Sistemas de Storage	Computadores pessoais/ dispositivos portáteis
	ESTRUTURAÇÃO	Banco de dados	Planilhas
COMPARTILHAMENTO	ACESSO	Acesso aberto/ distribuído	Obscuro ou protegido
	REUSO	Imediato/globalizado	Episódico/entre a equipe
INSTITUCIONALIZAÇÃO	FINANCIAMENTO	Fluxo contínuo/ Apoio internacional	Por projeto
	RECONHECIMENTO/ RECOMPENSA	SIM	NÃO


ORGANIZAÇÕES
UNIVERSIDADES
INSTITUTOS DE PESQUISA
AGÊNCIAS DE FOMENTO
BIBLIOTECAS, ARQUIVOS, MUSEUS
ORGANIZAÇOES VIRTUAIS;
COMUNIDADES
EXPERTISES
PESQUISADORES
CIENTISTAS DE DADOS
BIBLIOTECARIOS DE DADOS
ARQUIVISTAS
INSTRUMENTOS CIENTÍFICOS
TELECÓPIOS
SATÉLITES
COLISORES
SENSORES
CIBERINFRAESTRUTURA DE PESQUISA
DADOS
BASES DE DADOS
REPOSITÓROS
ACESSO
GESTÃO
CURADORIA
MINERAÇÃO
PRIVACIDADE
REDES
REDES DE PESQUISA/EDUCAÇÃO NACIONAIS E INTERNACIONAIS;
SEGURANÇA
RECURSOS COMPUTACIONAIS
SUPERCOMPUTADORES
NUVEM, GRID, CLUSTER;
VISUALIZAÇÃO;
CENTROS DE COMPUTAÇÃO
SOFTWARE
APLICAÇÕES;
DESENVOLVIMENTO
E SUPORTE



A CAUDA LONGA DA CIÊNCIA

A MAIORIA DAS COLEÇÕES DE DADOS PRODUZIDAS PELA PESQUISA CIENTÍFICA É GERADO/COLETADO POR PEQUENOS LABORATÓRIOS E PESQUISADORES INDIVIDUALMENTE NAS UNIVERSIDADES E INSTITUTOS DE PESQUISA, QUE DESENVOLVEM UM GRANDE NÚMERO DE PROJETOS CIENTÍFICOS

DOMÍNOS ESPECÍFICOS

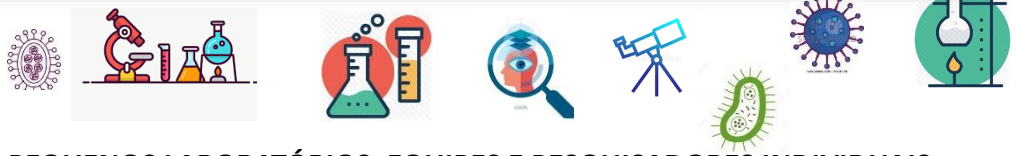
ASTRONOMIA
FISICA NUCLEAR
GENOMA
PROTEINA
SENSORIAMENTO REMOTO

Dados da Grande Ciência são fáceis de manipular, compreender e arquivar;
A Pequena Ciência é terrivelmente heterogênea e muito mais vasta e gera 2-3 vezes mais dados do que a Big Science (MacColl, 2010)

Volume dos dados

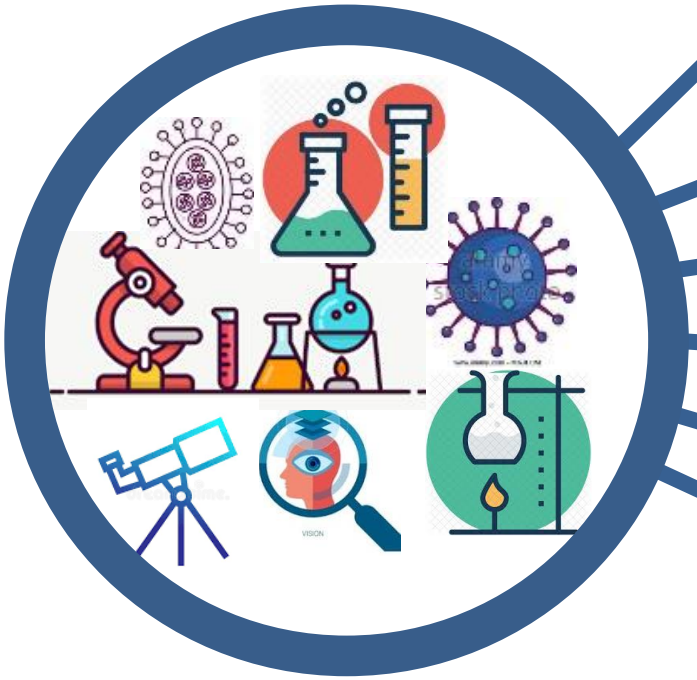
VÁRIOS DOMÍNIOS E INSTITUIÇÕES

Número de datasets



PEQUENOS LABORATÓRIOS, EQUIPES E PESQUISADORES INDIVIDUAIS

DADOS DA CAUDA LONGA:
HETEROGENEIDADE EM VÁRIAS DIMENSÕES



TAMANHO DAS COLEÇÕES

FORMATOS DOS DADOS

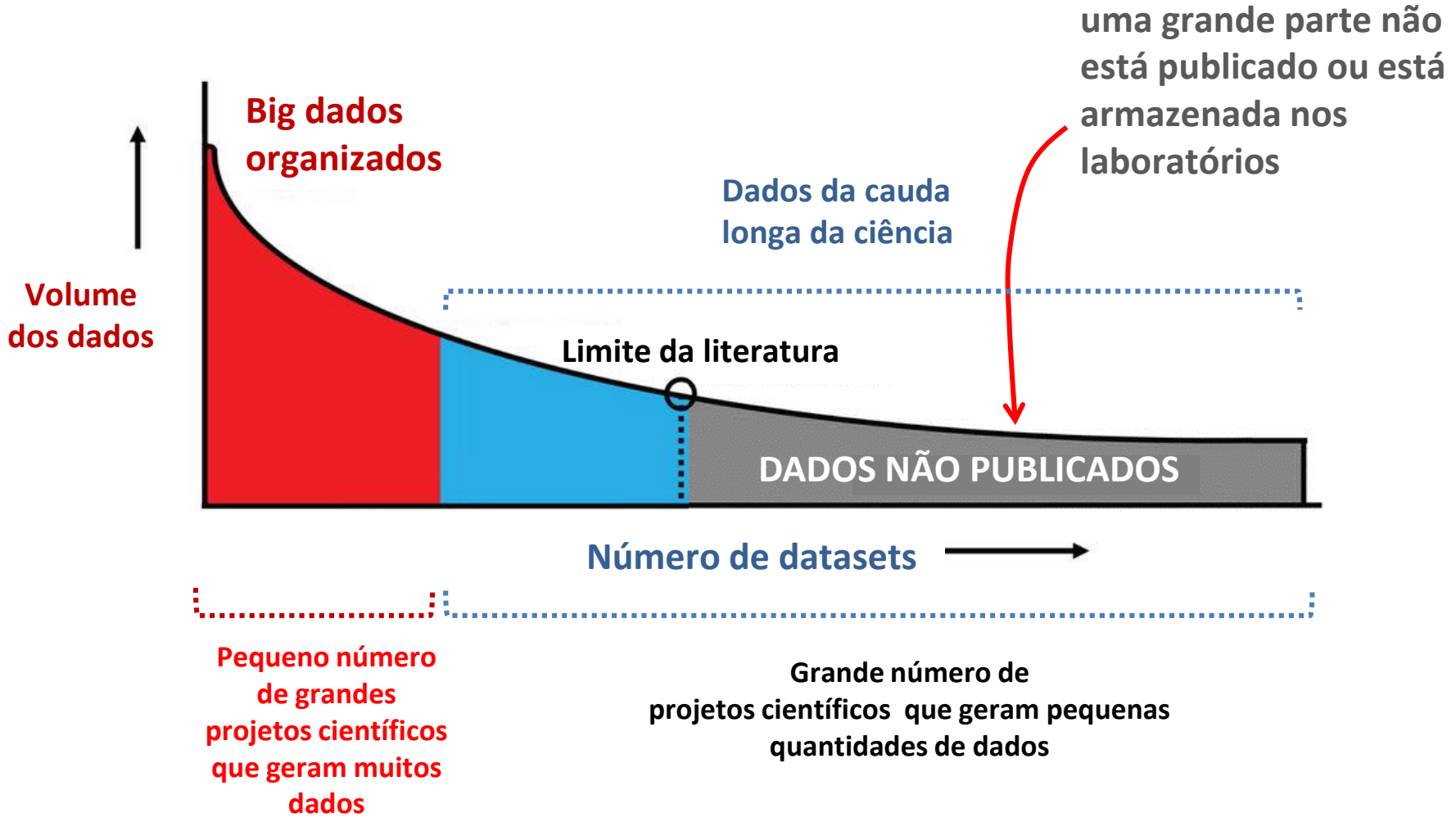
ESTRUTURA DOS DADOS

COMPLEXIDADE DOS DADOS

DOMÍNIOS DISCIPLINARES

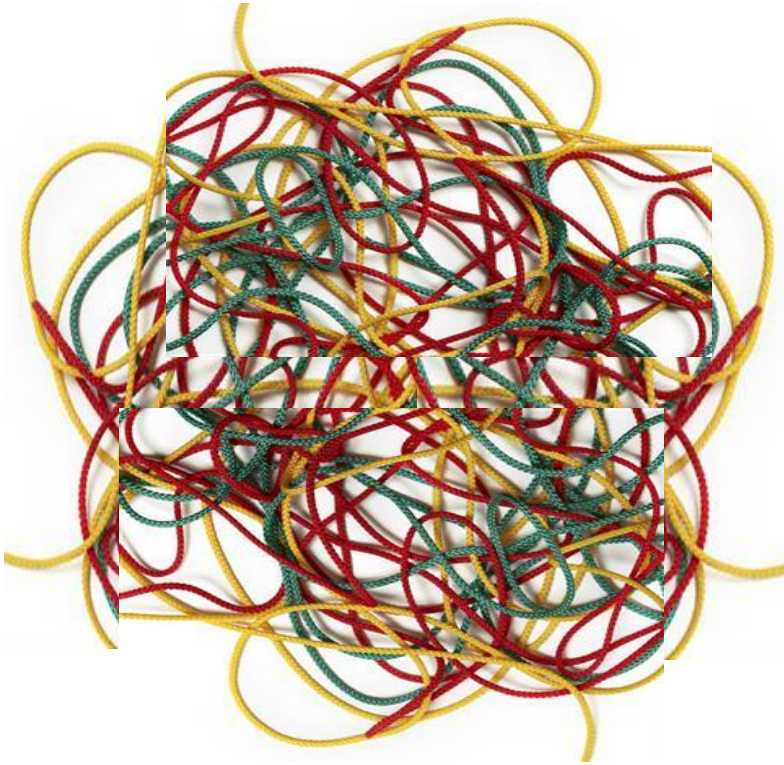
TECNOLOGIAS USADAS NA GERAÇÃO/COLETA

A CIEÑCIA INVISIÝEL DA CAUDA LONGA



Os dados gerados ou coletados em decorrência dos pequenos projetos de pesquisa são distribuídos por todos os domínios do conhecimento, das artes e humanidades até as áreas mais identificadas como os padrões da grande ciência como física e astronomia

“Parece mais provável que a ciência transformadora venha mais da cauda do que da cabeça (Heidorn, 2008)



modelos

patentes

Artigos científicos

Ciência disruptiva

Inovação

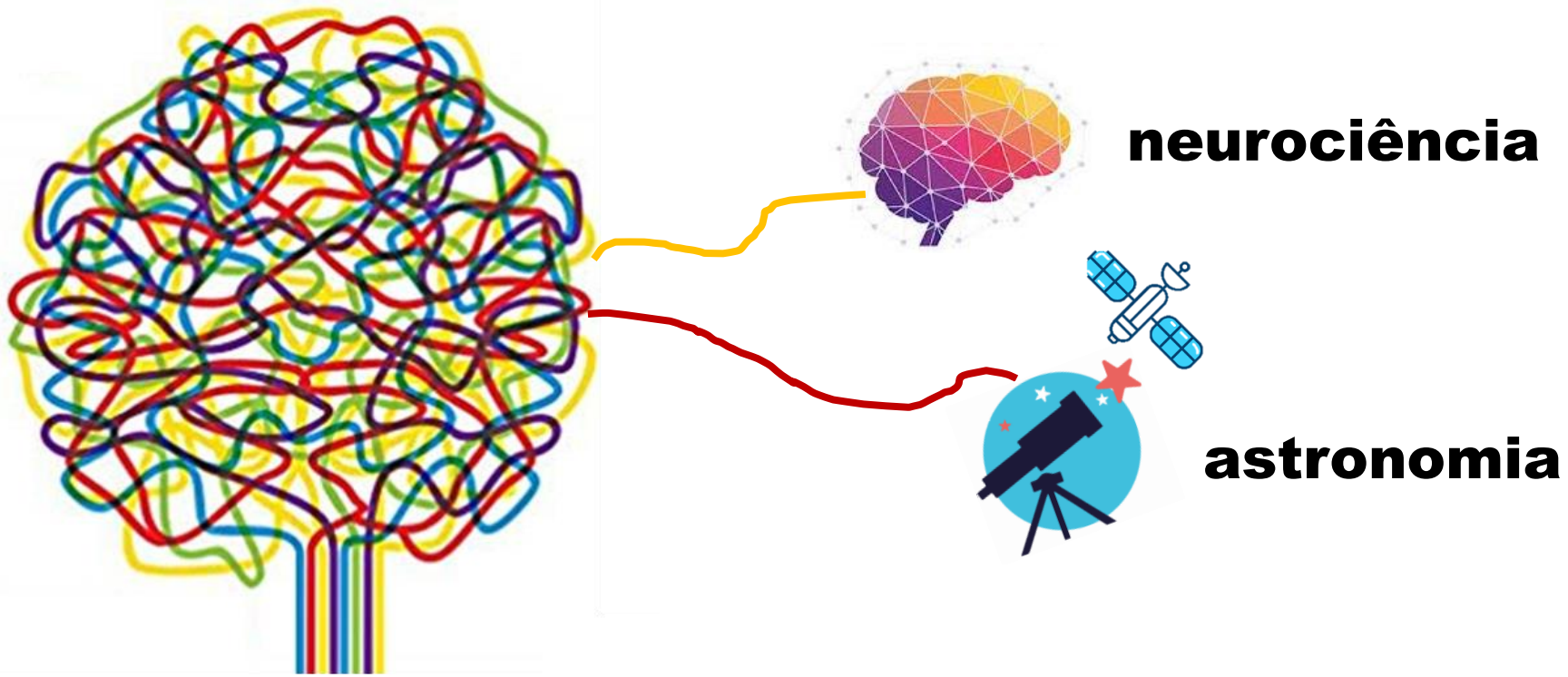
Independência criativa

Conhecimento multi e interdisciplinar

Autonomia

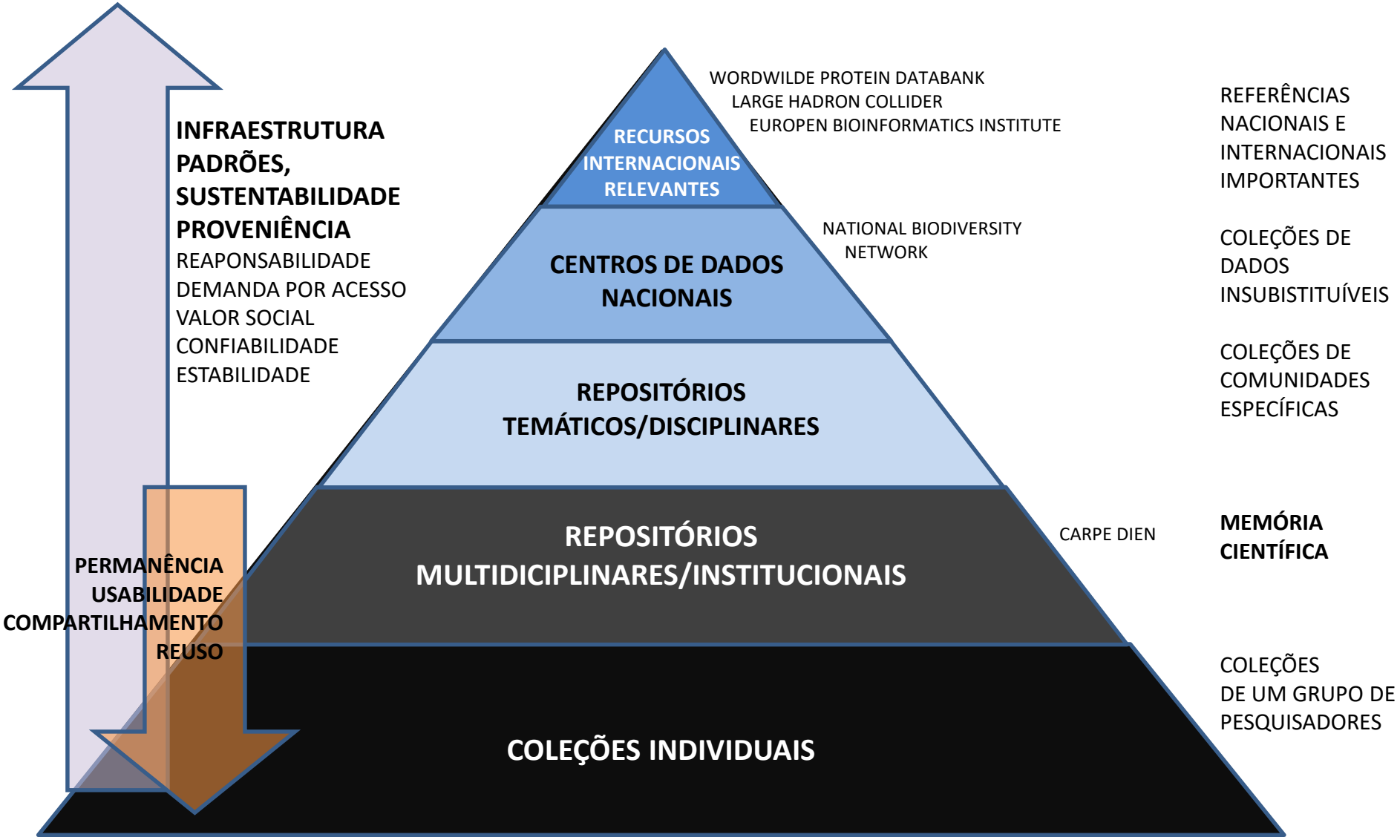
DIVERSIDADE DOS DADOS

Os dados da cauda longa, com **sua natureza heterogênea e diversificada**, devem se **integrar a homogeneidade da grande ciência** formando **uma ecologia ou diversidade de dados**. Isto por que nem sempre a grande ciência, definida por predicados homogêneos e estáveis **é o modelo mais adequado para algumas das áreas mais avançadas** e inovadoras da pesquisa científica. Na maioria das vezes, integrar dados formando uma diversidade de dados transversalmente rica, estabelece modelos eficientes de geração de conhecimento



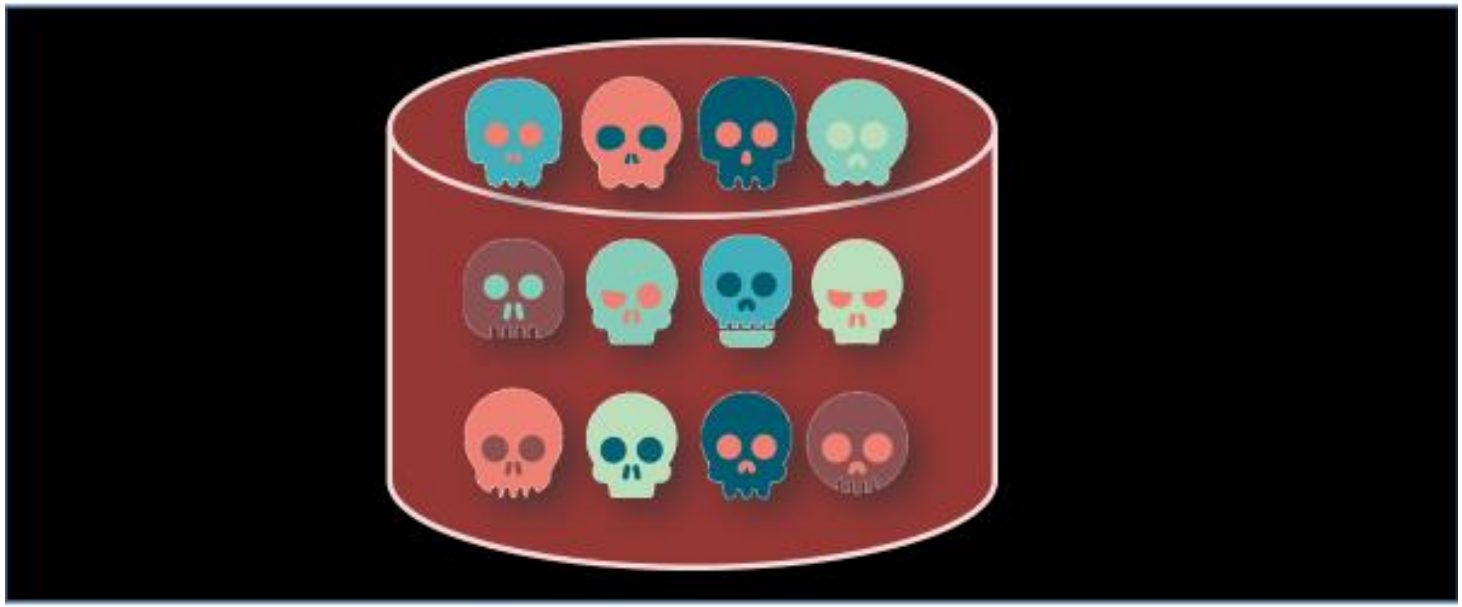
A **perspectiva sistêmica do espaço de dados** torna a integração desses ativos chave **para respostas a novas indagações da ciência**. Isso acontece especialmente ao vincular a estabilidade da grande ciência ao território de alto coeficiente de autonomia e independência da cauda longa, cujas condutas desafiadoras favorecem a inovação e a geração de conhecimentos multi e interdisciplinar.

PIRÂMIDE DE GESTÃO DE DADOS



O sucesso dos novos serviços de informação para a pesquisa está relacionado à sua capacidade de dar apoio às **práticas e culturas** das comunidades científicas da instituição.

REPOSITÓRIO DE DADOS



CEMITÉRIO DE DADOS ?

Por que os pesquisadores da cauda longa não compartilham seus dados?

A maioria dos pesquisadores concordam em tese com os princípios de compartilhamento e reuso preconizados pela ciência aberta, mas relutam em compartilhar os seus próprios dados como parte do fluxo de pesquisa , e o fazem mais como exceção do que como regra .

MOTIVOS PARA O PESQUISADOR NÃO COMPARTILHAR

Restrições culturais, DISCIPLINARES e institucionais

x

INTERESSES ECONÔMICOS (patentes, acordos comerciais, etc)

RESULTADOS NEGATIVOS, hipóteses não confirmadas

CUSTO do tratamento dos dados (limpeza, catalogação, formatos, etc.)

Perda da VANTAGEM COMPETITIVA de publicar mais baseado nos dados

Preocupação dos dados serem ERRONEAMENTE INTERPRETADOS por outros pesquisadores

Dificuldade de garantir a PRIVACIDADE dos dados

+50%

DOS ACHADOS NÃO FORAM PUBLICADOS

LADO OCULTO DOS DADOS

DADOS NÃO PUBLICADOS

INDIVIDUAIS
Personalidade e **atitudes** do pesquisador em relação aos seus dados.

DISCIPLINARES
Cultura do compartilhamento do domínio

ORGANIZACIONAIS
Obstáculos inerentes às instituições de pesquisa em relação à gestão de dados

POLITICAS
Condicionantes políticas voltadas para **recompensa, financiamento e sustentabilidade** de longo prazo

INFRAESTRUTURAIS
Aportes tecnológicos disponíveis

O COMPRTILHAMENTO PODE REVELAR VALORES IMPORTANTES OCULTOS NESSES DADOS

AGÊNCIAS FINANCIADORAS DE PESQUISA

PLANOS DE COMPARTILHAMENTO DE DADOS

POLÍTICAS MANDATÓRIAS

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

FAPESP

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).

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 .

nature.com

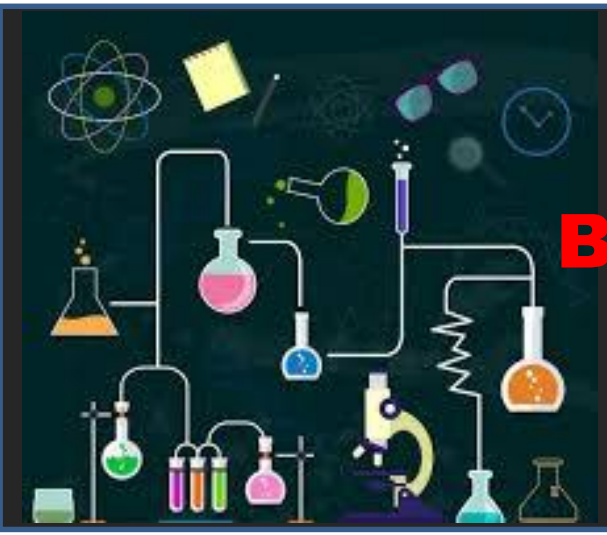
The world's best science and medicine on your desktop

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.

BARREIRAS DISCIPLINARES

Para algumas disciplinas o compartilhamento é determinante para a geração de conhecimento, para outros é somente uma ação entre colegas.



BARREIRAS ORGANIZACIONAIS

POLÍTICA DE GESTÃO DE DADOS
SERVIÇOS
BALCÃO DE REFERÊNCIA PARA DADOS
TREINAMENTO
FERRAMENTAS DE SOFTWARE
EXPERTISE EM GESTÃO DE DADOS

POLÍTICAS DE DADOS

DIRETRIZES INTERNACIONAIS

PADRÕES INTERNACIONAIS

TECNOLOGIAS ABERTAS

EDITORAS CIENTÍFICAS

NACIONAL GOVERNAMENTAL

Política de ct & i
Conformidade legal/responsabilidades
Transparência
Proteção à propriedade intelectual
Ética
Prioridades estratégicas

FOMENTO À PESQUISA

Políticas mandatórias
Financiamento
Sustentabilidade
Prioridades

INSTITUCIONAL

Aquisição/desenvolvimento de coleções
Treinamento
Segurança/armazenamento
Preservação
Boas práticas/qualidade
Infraestrutura tecnológica
Arquivamento
Publicação
Licenças

REPOSITÓRIOS

depósito/acesso
tipos de dados
interoperabilidade
formatos
Identificadores persistentes
curadoria
serviços
metadados/documentação
tempo de embargo

POLÍTICA DE COMUNIDADES/DISCIPLINAS/COLEÇÕES

Como a ciência é um **empreendimento humano**, argumentos a favor e contra o compartilhamento de dados frequentemente focam menos nos benefícios percebidos para a ciência como um todo e **mais nos efeitos sobre o pesquisador individualmente** (FERGUSON et al, 2014)

BARREIRAS INDIVIDUAIS

Oportunismo de outros pesquisadores
quando eles disponibilizam os seus materiais de pesquisa nas fases preliminares do processo de pesquisa, expondo-os a **abusos nos seus direitos intelectuais**;

Ser "furado"
ou seja, de ter publicações baseada nos seus dados lançadas em primeira mão por outros autores.

Explorar mais os dados
publicando o máximo possível de artigos baseados nesses dados, já que esse é o **critério academicamente mais valorizado**

Reanálises equivocadas
Dados de má qualidade e análises por não especialistas

Erros nos dados ou nas análises
temor que outros pesquisadores descubram erros nos dados ou questionam a validade das análises.

Tempo, esforço e recurso
Para que as coleções de dados possam ser reusados de ser limpas; organizadas, documentadas, anonimizadas e descritas por metadados que evidenciem os instrumentos e métodos usados para obtê-las e, finalmente, publicadas em bases de dados/repositórios

Falta de reconhecimento pelos sistemas de recompensa por organizar os dados

Falta de conhecimento das tecnologias para o compartilhamento

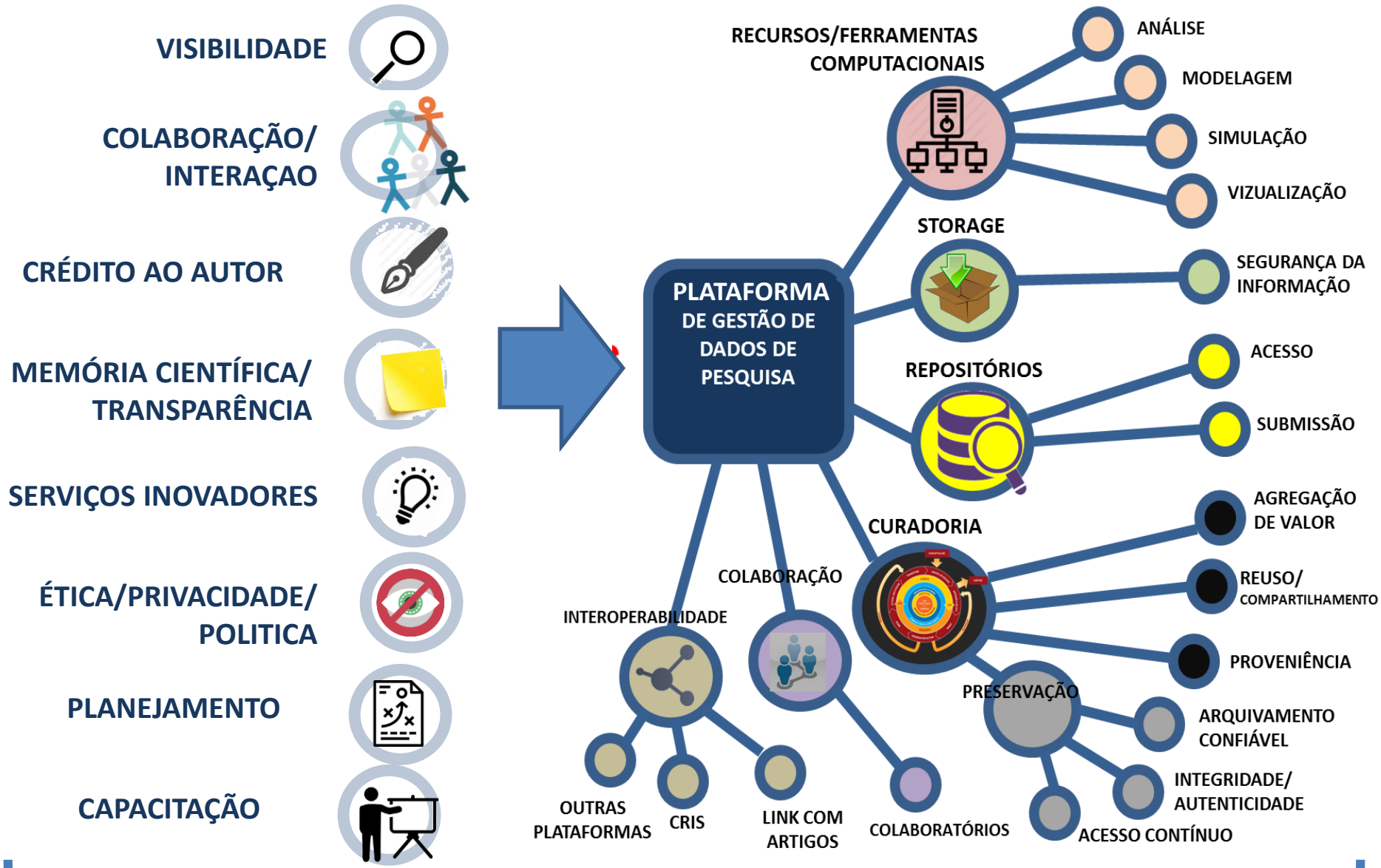
Resultados de experimentos que não deram certo

Problemas éticos, de privacidade e legais
"A maior barreira para o compartilhamento de dados de pesquisa são os temores dos pesquisadores **em relação às questões legais e o mau uso de seus dados**"

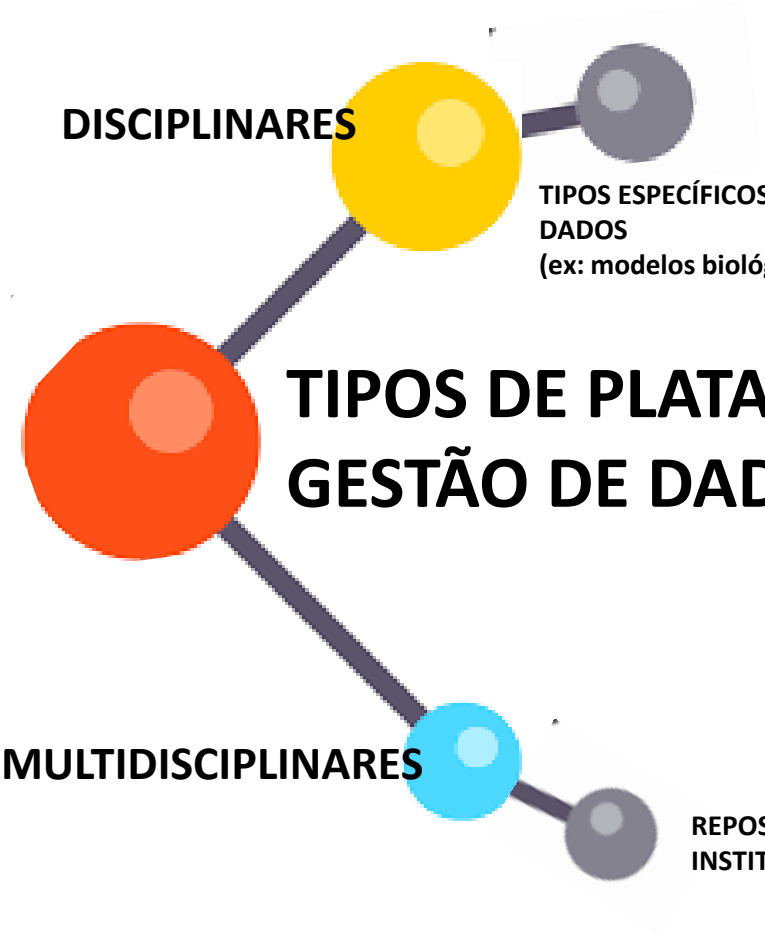
Baixo impacto na carreira do pesquisador.

Interesse econômico

PLATAFORMA DE GESTÃO DE DADOS DE PESQUISA



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

POR QUE A GESTÃO DE DADOS DE PESQUISA NECESSITA DE NOVAS ABORDAGENS?

AFINAL, O QUE E DADO DE PESQUISA?

> Informação é um conceito complexo com centenas de definições [...]. Dado [por sua vez] é um conceito simples com poucas definições, porém sujeito a muitas e diferentes interpretações

CRISTINE BORGMAN (2007, P.9)

O que dificulta atribuir uma definição consensual ao dado de pesquisa é o fato idiossincrático que ele pode ser muitas coisas diferentes para pessoas e circunstancias diferentes. Isto acontece porque dado de pesquisa é dependente de interpretação.

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

CRITICOS

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

DINÂMICOS

ESTÁVEIS

EXPERIMENTAIS

OBSERVACIONAIS

ÊFEMEROS

TERCIÁRIOS

COMPUTACIONAIS

SUJOS

PRIMÁRIOS

DERIVADOS

LIMPOS

TIPOS

TEXTO

SIMULAÇÃO

VIDEO

SOFTWARE

NÚMEROS

GRÁFICOS

FÓRMULA

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**.

FLUXO DOS DADOS

A MAIOR PARTE DOS DADOS NÃO É DIRETAMENTE ÚTIL NO MOMENTO EM QUE COLETADA

INSTRUMENTO CIENTÍFICO

DADOS BRUTOS

PROCESSAMENTO DOS DADOS

Selecionar subset
Mesclar mútiplos datasets
Conversão
Normalização
Limpeza dos dados

COMPUTAÇÃO EM NUVEM
COMPUTAÇÃO EM GRADE

ANÁLISE DOS DADOS

Estatísticas
Simulação
Plotagem
Visualização
Modelos
Algoritmos
Publicações

FLUXO DOS DADOS

UM EXEMPLO DE PROCESSAMENTO DE DADOS BRUTOS

pluviômetro

```
0711070500276000
0711070600276000
0711070700277003
0711070800282017
0711070900285000
0711071000293000
0711071100301000
0711071200304000
```

dados brutos

Date	time	air temp C	precip mm
11-Jul-07	5:00	27.6	000
11-Jul-07	6:00	27.6	000
11-Jul-07	7:00	27.7	003
11-Jul-07	8:00	28.2	017
11-Jul-07	9:00	28.5	000
11-Jul-07	10:00	29.3	000
11-Jul-07	11:00	30.1	000
11-Jul-07	12:00	30.4	000

Recreated from Michener & Brunt (2000)

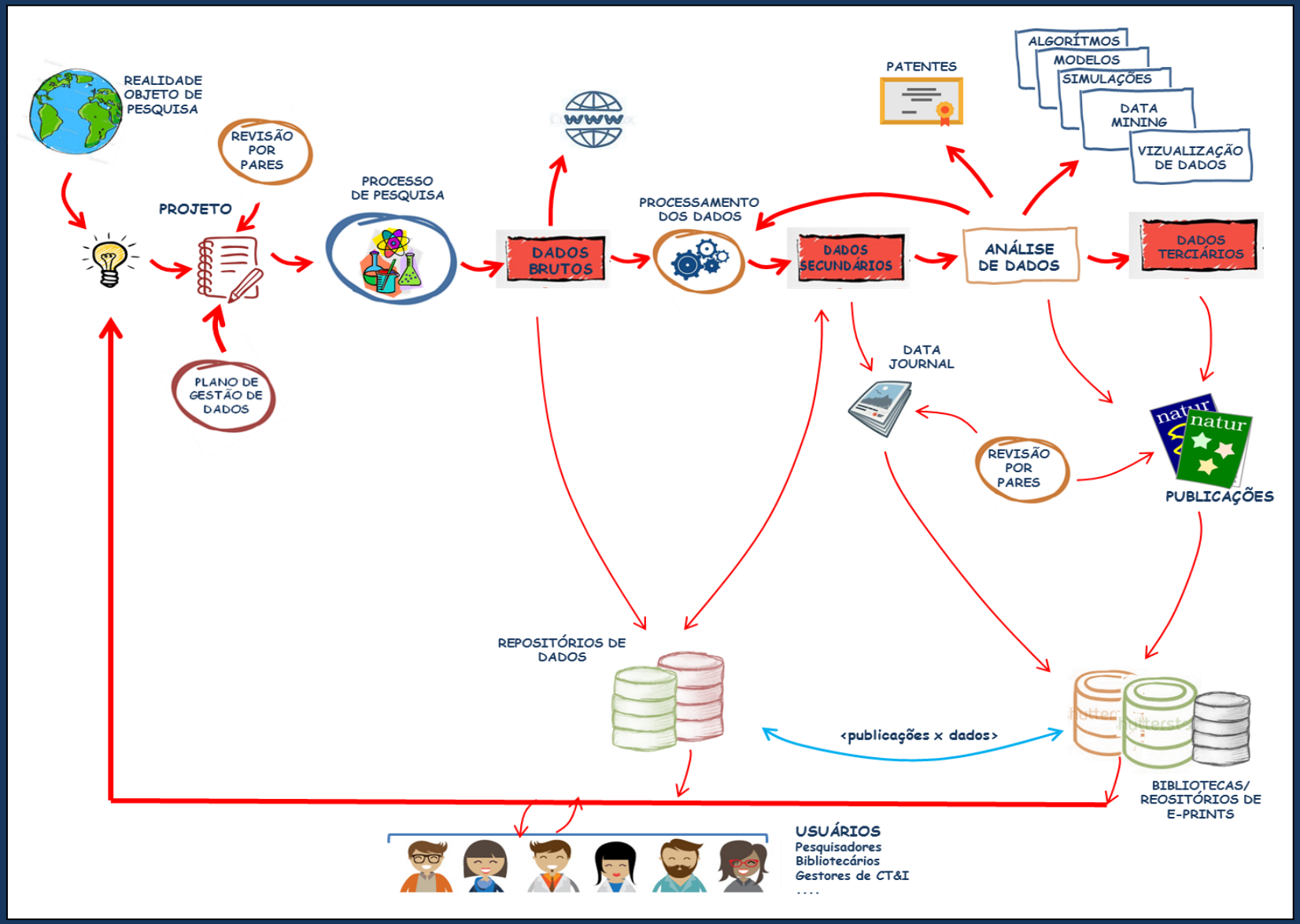
dados processados

É NECESSÁRIO COMPARTILHAR MUITO MAIS DO QUE OS DADOS FINAIS PARA A PESQUISA SEJA REPRODUTÍVEL

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

É MAIS DIFÍCIL VALIDAR OS DADOS SE AS INFORMAÇÕES ENTRE A IDEIA INICIAL E RESULTADO FINAL ESTÃO FALTANDO

UM CICLO DE VIDA PARA OS DADOS DE PESQUISA



Experimento > análise > publicação

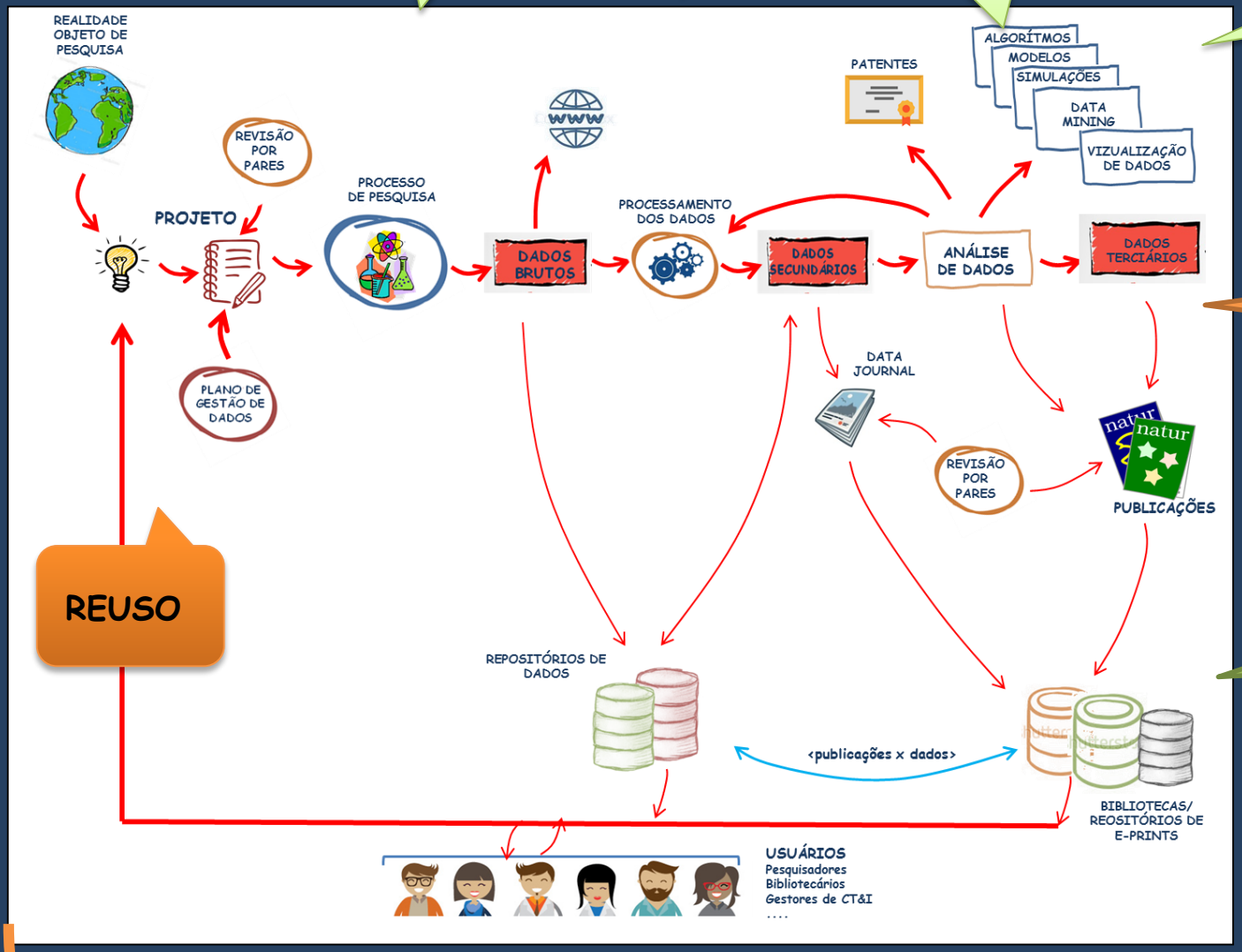
Experimento > organização de dados > análise > publicação

CAPTURA DE DADOS

LIMPEZA DOS DADOS

ANÁLISES E RESULTADOS

PROCESSOS COMPUTACIONAIS



METADADOS, DOCUMENTAÇÃO, VERSIONAMENTO

INTEGRAÇÃO COM O SISTEMA DE PUBLICAÇÃO

REUSO

ARQUIVAMENTO PRESERVAÇÃO

INTEROPERABILIDAE

DESCOBERTA & ACESSO

POLÍTICAS – SUSTENTABILIDADE – CONFORMIDADE LEGAL E ÉTICA

DEFINIÇÃO DE POLÍTICAS

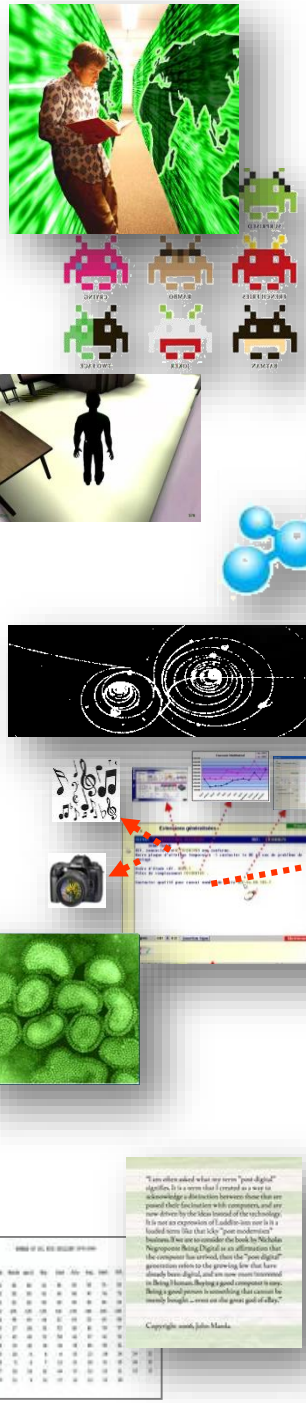
OBJETOS DIGITAIS COMPLEXOS

Texto e números não contam toda história

DADOS DE PESQUISA SÃO **OBJETOS COMPLEXOS, DIVERSIFICADOS E HETEROGÊNEOS.**

OS OBJETIVOS E OS MÉTODOS USADOS PARA PRODUZI-LOS VARIAM ENORMEMENTE DE ACORDO COM OS CAMPOS CIENTÍFICOS, **ASSIM COMO OS CRITÉRIOS PARA COMPARTILHÁ-LOS,**

NÍVEIS DE ABSTRAÇÃO



REALIDADE VIRTUAL

GAMES

SIMULAÇÕES

MODELOS EM 3D

ESTRUTURAS QUÍMICAS

SOFTWARE

WEBSITE/MULTIMÍDIA

VIDEOS

FOTOS

GRÁFICOS

ESPECIFICAÇÕES

ENTREVISTAS

FORMÚLAS

TABELAS

ANOTAÇÕES

DADOS NUMÉRICOS

dispositivos de imersão e interativas

apresentações sensoriais

imagem em movimento

imagens

sons

documentos

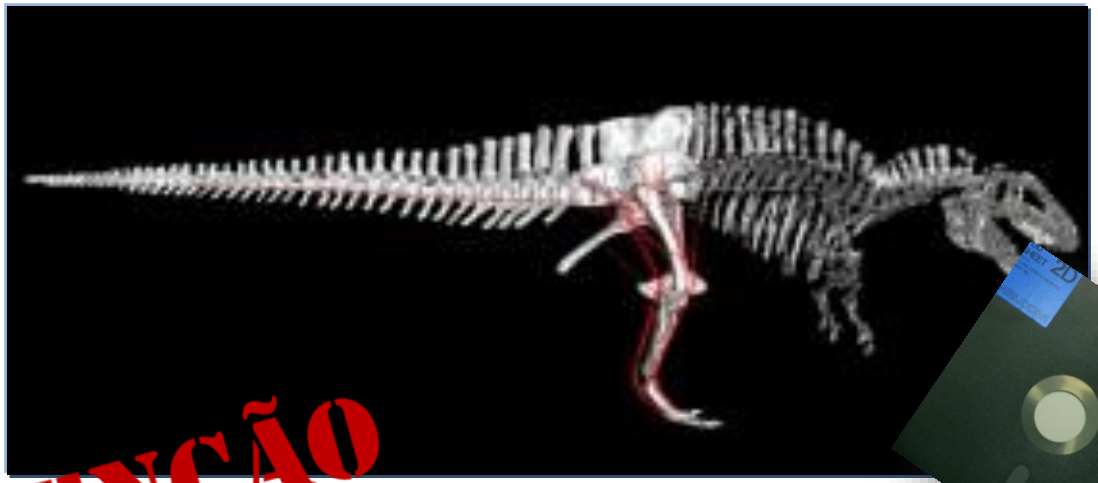
letras

símbolos

números

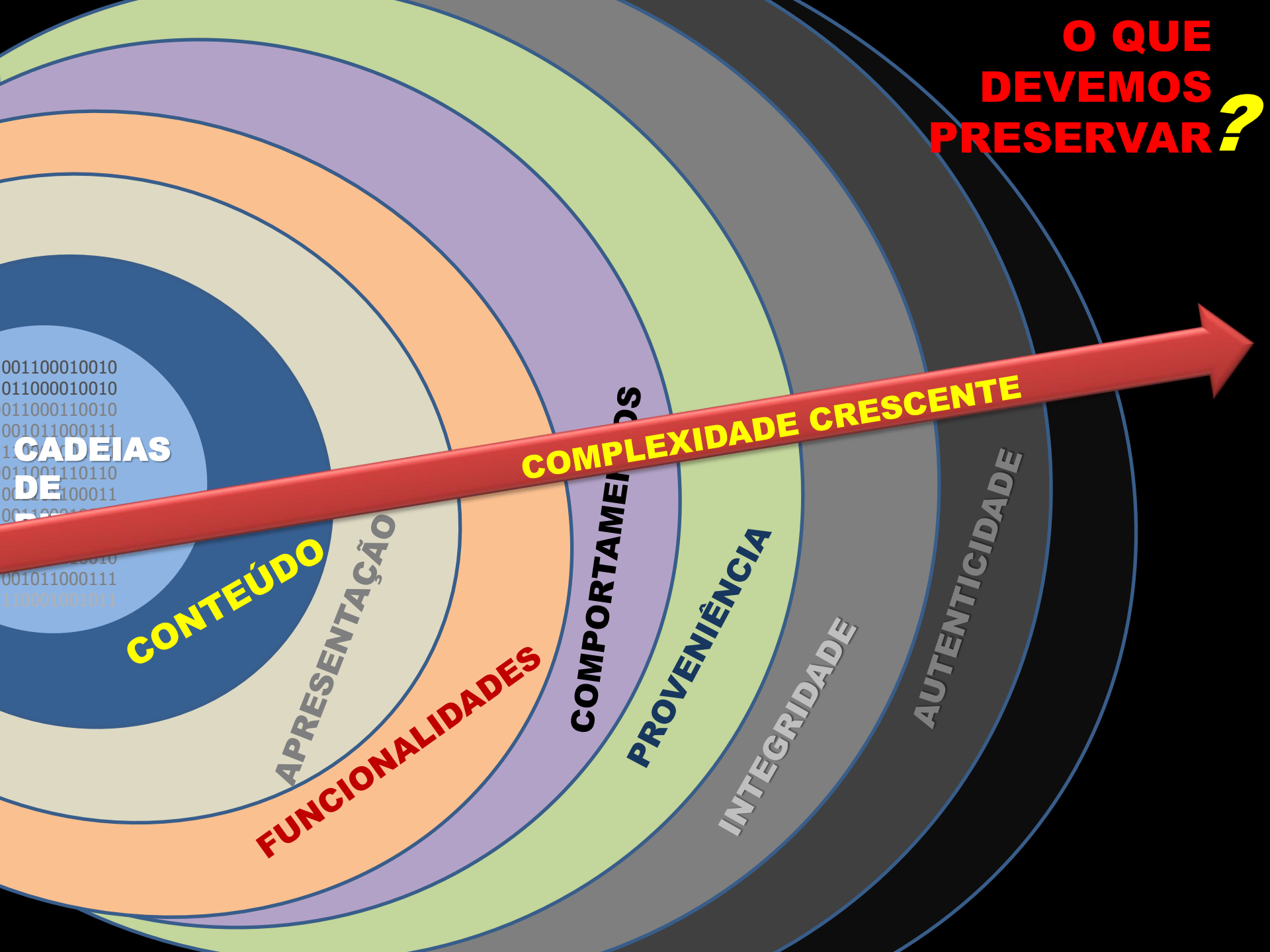
OBJETOS DIGITAIS COMPLEXOS

A INFORMAÇÃO DIGITAL NÃO SOBREVIVE INERCIALMENTE



INTENÇÃO

A mesma tecnologia que muda a pesquisa científica coloca os dados gerados em risco e nos impõe o desafio estratégico, gerencial e político de criar, arquivar, preservar e tornar disponível esses dados


O QUE DEVEMOS PRESERVAR?
CADEIAS DE
CONTEÚDO
APRESENTAÇÃO
FUNCIONALIDADES
COMPORTAMENTOS
PROVENIÊNCIA
INTEGRIDADE
AUTENTICIDADE
COMPLEXIDADE CRESCENTE

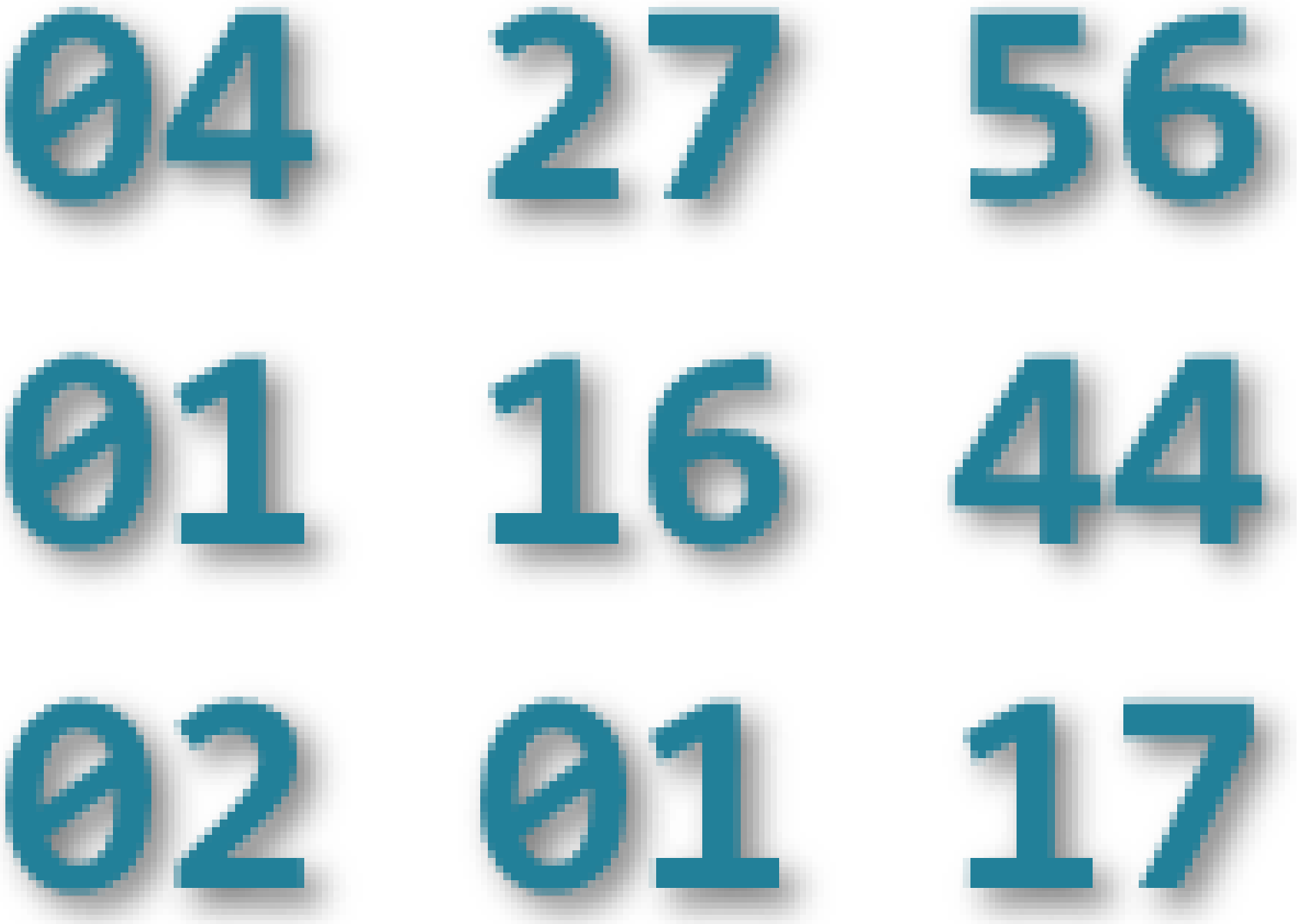
KEEP CALM

E ACREDITE

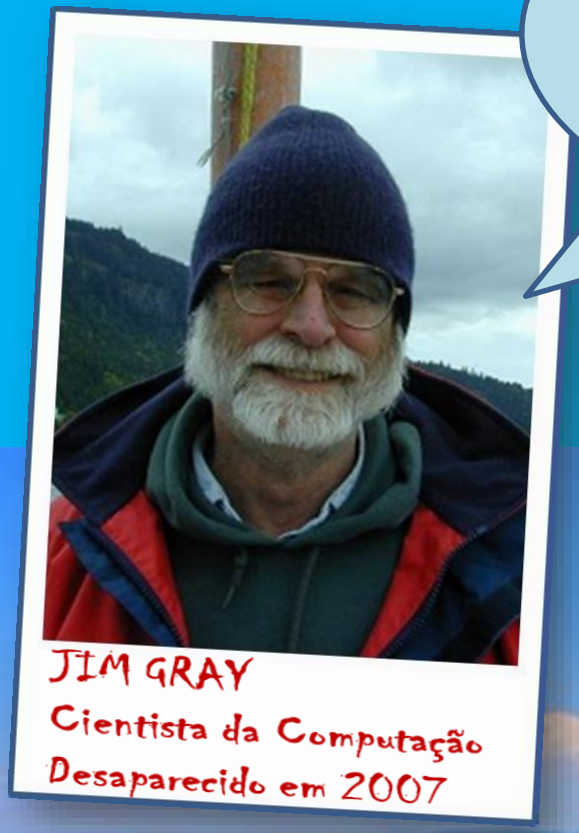
NOS REPOSITÓRIOS DIGITAIS CONFIÁVEIS

CONTEXTO e ESTRUTURA DOS DADOS

SIGNIFICADO


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 **!!!**

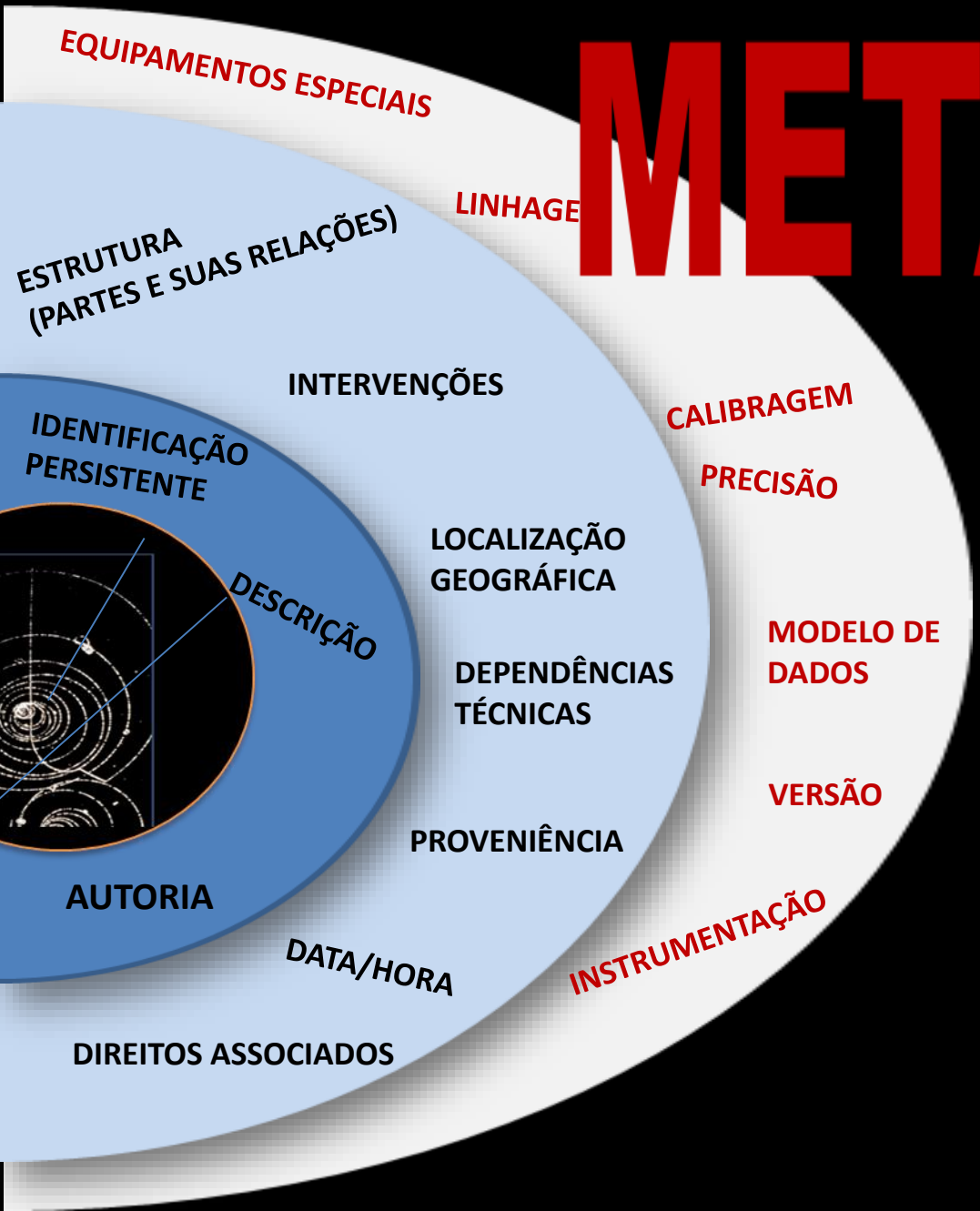
Para entender os dados os usuários futuros necessitam de metadados, caso contrário eles não saberão os detalhes de como os dados foram **obtidos e preparados** : 1) como os **instrumentos foram projetados e construídos**; 2) **quando, onde e como** os dados foram coletados; e 3) e não terão **uma descrição dos processos que levaram aos dados derivados**, que são tipicamente usados para analises científicas de dados.

Gray, 2002

SIGNIFICADO
ESTRUTURA
IDENTIFICAÇÃO
CONTEXTO

PROVENIÊNCIA

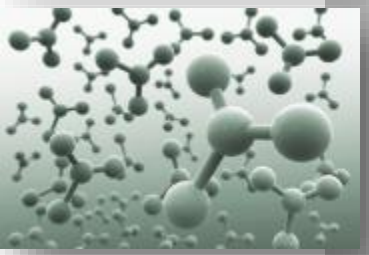
METADADOS



Os metadados têm um forte impacto na capacidade dos dados de pesquisa de transmitir conhecimentos e poder ser interpretados e reusados agora e no futuro

DESCRITIVOS
ADMINISTRATIVOS
TÉCNICOS
ESTRUTURAIS
PRESERVAÇÃO

DISCIPLINARES



METADADOS DISCIPLINARES

GENOME METADA

ECOLOGICAL METADATA LANGUAGE

ASTRONOMY VIZUALIZATION METADATA

DDI-DATA DOCUMENTATION INITIATIVE

AGRICULTURAL METADATA ELEMENT SET

DARWIN CORE

http://www.dcc.ac.uk/drupal/resources/metadata-standards

INFORMAÇÕES DE REPRESENTAÇÃO

METADADOS

- DISCIPLINARES
- GERAIS
  - DESCRITIVOS
  - ADMINISTRATIVOS
  - ESTRUTURAIS
  - TÉCNICOS
  - PRESERVAÇÃO

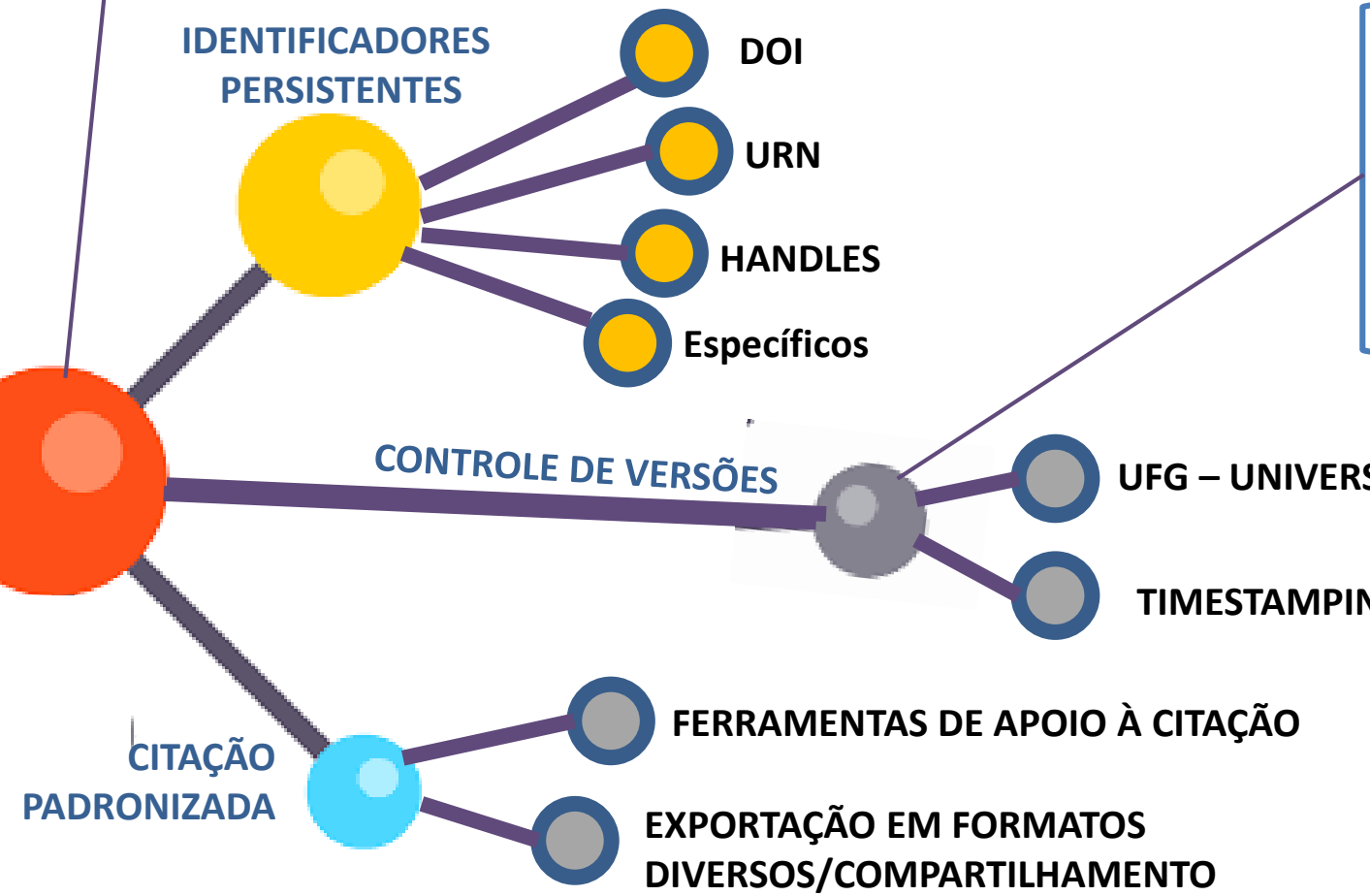
DOCUMENTAÇÃO

- CADERNO DE LABORATÓRIO
- CADERNO DE CAMPO
- ROTEIRO DE ENTREVISTA
- PROJETO
- ARQUIVO “LEIA-ME”
- ANOTAÇÕES
- DATA PAPER

IDENTIFICAÇÃO

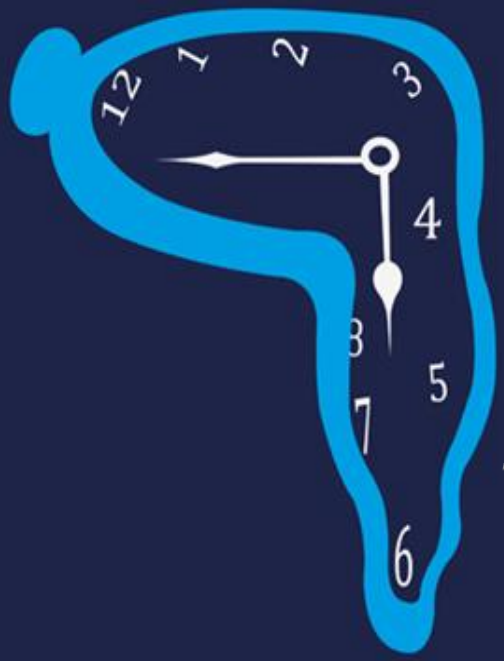
REFERÊNCIA

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*.



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

VERSION 01
VERSION 02
VERSION 03
VERSION 04
VERSION 05



AS COLEÇÕES DE DADOS PODEM EVOLUIR COM O TEMPO

- PESQUISAS EM ANDAMENTO
- INCORPORAÇÃO DE NOVOS DADOS
- REPROCESSAMENTO DA COLEÇÃO
- CORREÇÕES
- APRIMORAMENTO DOS MÉTODOS DE GERAÇÃO/COLETA

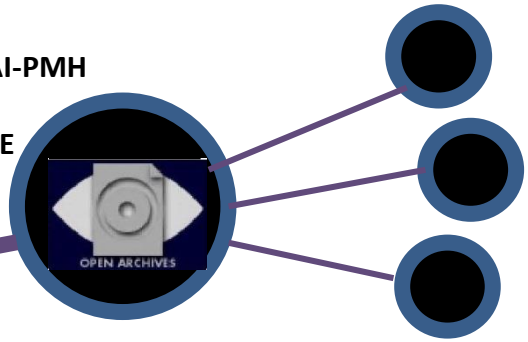
IDENTIFICADAS/CITADAS

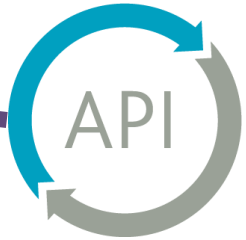
- **UNF – UNIVERSAL NUMERIC FINGERPRINT**
- **TME STAMP**

INTEROPERABILIDADE

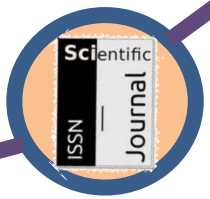
A INTEROPERABILIDADE DE SISTEMAS É COMPREENDIDA AQUI COMO A CAPACIDADE DAS PLATAFORMAS DE REPOSITÓRIOS DE DADOS INTERCAMBIAREM INFORMAÇÕES – **DADOS E METADADOS** - COM SISTEMAS EXTERNOS DE FORMA HARMÔNICA E INTEGRADA E COM PROPÓSITOS ESPECÍFICOS.

PROTOCOLO OAI-PMH PARA COLETA AUTOMÁTICA DE METADADOS





PROGRAMAS PARA EXPOR METADADOS

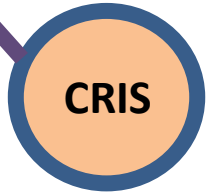


SISTEMAS DE PUBLICAÇÕES DE PERIÓDICOS

INTEGRAÇÃO COM OUTROS SISTEMAS

LINKING - USO DO LINKED DATA

EXPORTAÇÃO DE METADADOS - DC, MARC-XML, METS...



SISTEMA DE GESTÃO DE PESQUISA

SISTEMA DE REPOSITÓRIO CONFIÁVEL

ARCHIVEMÁTICA

Nesse contexto a compreensão de três tipos de dados se torna essencial:

DADOS PESSOAIS São dados relacionados a indivíduos vivos, que podem ser identificados a partir desses dados ou a partir desses dados combinados com outras informações.

DADOS CONFIDENCIAIS São dados que não estão em domínio público tais como informações sobre negócios, lucros, saúde, detalhes médicos e opiniões políticas, entregues em confiança ou que duas partes concordam em mantê-los confidenciais, isto é, secretos.

DADOS PESSOAIS SENSÍVEIS São dados sobre raça, origem étnica, opinião política, religião ou crenças similares, filiação sindical, doença física ou mental, vida sexual, etc.

CURADORIA DIGITAL

imensas e sempre crescentes quantidades de conteúdos digitais

objetos complexos e heterogêneos, que dependem de tecnologias específicas e pouco duradouras,

grande diversidade de contextos organizacionais em que a curadoria de conteúdos digitais ocorre;

uma audiência que pode ser indefinida e localizada no futuro

A **CURADORIA DIGITAL** difere, em **termos de significado e amplitude conceitual**, da **CURADORIA** como ela vem sendo compreendida ao longo do tempo**!**

Entretanto, a **curadoria digital** mostra alguma continuidade com as **praticas tradicionais de curadoria!**

Independente de uma coleção **ser constituídas de objetos físicos ou digitais** – ou seja, de átomos e moléculas ou de bits e bytes - um curador deve **avaliar seu valor** e **relevância para a comunidade de usuários reais e potenciais**; determinar a **necessidade de preservação**; **documentar a origem e autenticidade**; **descrever, registrar e catalogar seu conteúdo; providenciar armazenamento e preservação** a longo prazo; e proporcionar um **meio de acesso e uso para os conteúdos** (NRC, 2015).

DESCONSTRUINDO O CONCEITO DE CURADORIA DE DADOS DE PESQUISA

AÇÕES GERENCIAIS, TECNOLÓGICAS E POLÍTICAS

NECESSÁRIAS PARA MANTER OS DADOS POR TODO O SEU CICLO DE VIDA – DESDE A SUA CRIAÇÃO -VISANDO O USO CORRENTE E FUTURO

QUE PRESSUPÕE ADICIONAR VALOR

ORGANIZAÇÕES EM COLEÇÕES
DOCUMENTAÇÃO
ATRIBUIÇÃO DE METADADOS
IDENTIFICAÇÃO
ARQUIVAMENTO
PRESERVAÇÃO
SEGURANÇA FÍSICA
AVALIAÇÃO (AUTENTICAÇÃO E VERIFICAÇÃO)
CONTROLE DE QUALIDADE
ANOTAÇÃO
LINKS

PARA GARANTIR

INTELIGENTEMENTE
ABERTOS
COMPREENSÍVEIS
LONGEVOS
DISPONÍVEIS
RECUPERÁVEIS
ACESSÍVEIS
AVALIÁVEIS
(PROVENIÊNCIA/ INTEGRIDADE/ QUALIDADE)
CONFORMIDADE LEGAL E ÉTICA
PADRONIZADOS
INTEROPERÁVEIS

COM OBJETIVO FINAL

USABILIDADE/REUSO
REPRODUTIBILIDADE
INTERDISCIPLINARIDADE
INPUT PARA NOVAS PESQUISAS
ENSINO DAS CIÊNCIAS
MEMÓRIA ACADÊMICA
VALIDAÇÃO DA PESQUISA

VOLTADO PARA UM PÚBLICO-ALVO

ALINHADO COM O FLUXO DE PESQUISA

REUSO EM OUTROS CONTEXTOS

ARQUIVOLOGIA

Centenas de **diários de bordo** digitalizados, registrando viagens marítimas de três séculos

DATA MINING

se tornam uma base de dados rica sobre a **fauna, flora, corrente e ventos oceânicos**

CLIMATOLOGIA

Cientistas reconstroem a história dos **sistemas dinâmicos da Terra** e melhoram as **projeções sobre o futuro do clima**

REUSO DE DADOS DE PESQUISA

EM OUTROS CONTEXTOS

"**Gerar, capturar e gerenciar coleções de dados** para uso corrente e futuro **não é um desafio trivial**. **As atividades que são locais e tácitas se tornam globais e explícitas**".

Tornar um conteúdo que foi criado para uma audiência útil para outra é um problema complexo, porque cada disciplina tem seu próprio **vocabulário, estrutura de dados** e **práticas de pesquisa** e **formulam questões** de forma distintas usando sua própria **terminologia**. Isto coloca um desafio importante para os **serviços de curadoria** que é criar descrições e representações, ferramentas e serviços que tornem viável o compartilhamento entre diferentes audiências (BORGMAN,2007).

PRINCÍPIO FAIR

ENCONTRÁVEL:
Fácil de achar por humanos e computadores por meio de metadados que facilitem a busca por datsets específicos.

ACESSÍVEL:
Armazenado por longo prazo de forma que ele pode ser facilmente acessado e/ou baixado com licenças e condições de acesso bem definidas (acesso aberto quando possível)

INTEROPERÁVEL
Pronto para combinar com outros dados por seres humanos ou por computadores

REUSÁVEL
Pronto para ser usado para pesquisas futuras, e para ser processado usando métodos computacionais.


DADO DE PESQUISA MANEIRO!
IDENTIFICADO
CITADO
VISÍVEL
LOCALIZADO
RECUPERADO
ACESSADO
INTERPRETADO
CONTEXTUALIZADO
AVALIADO
PROVENIÊNCIA
REUSADO
COMPARTILHADO
ON-LINE
ANOTADO
ATIVA COLABORAÇÃO
INTEROPERÁVEL
LINKADO COM
PUBLICAÇÃO
LICENÇA APROPRIADA
CONSIDERA PRIVACIDADE/ÉTICA
ARQUIVADO
PRESERVADO

CONTORNANDO A INVISIBILIDADE DA CAUDA LONGA


RECONHECIMENTO/RECOMPENSA
CITAÇÃO/FATOR DE IMPACTO/PUBLICAÇÃO
INCENTIVOS
POLÍTICAS
EXIGÊNCIAS PARA FINANCIAMENTO DE PROJETOS
EXIGÊNCIAS DO EDITORES
PESQUISADORS COMO PRODUTORES E USUÁRIOS
CAPACITAÇÃO
BIBLIOTECARIOS DE DADOS
CIENTISTAS DE DADOS
INFRAESTRUTURA
ÊNFASE DISCIPLINAR
FINANCIAMENTO
ORIENTAÇÃO POR PROJETOS/FINANCIMENTO PERMANENTE
CUSTO-BENEFÍCIO (

PUBLICAÇÃO DE DADOS

PUBLICAÇÃO EM REPOSITÓRIO DE DADOS DISCIPLINAR/TEMÁTICO

PUBLICAÇÃO EM REPOSITÓRIO DE DADOS MULTIDISCIPLINAR

PUBLICAÇÃO EM DATA JOURNAL

PUBLICAÇÃO EM PERIÓDICOS COMO MATERIAL SUPLEMENTAR

PUBLICAÇÃO EM REPOSITÓRIO INSTITUCIONAL

WEBSITE DO PROJETO OU DA INSTITUIÇÃO

PEN DRIVE
NOTEBOOK

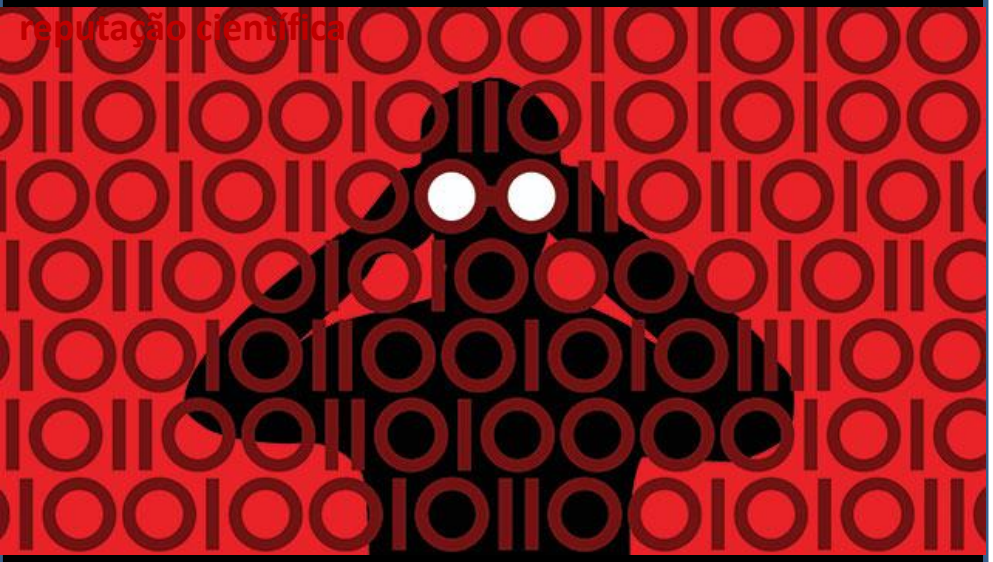
VIVIBILIDADE

REUSO
COMPARTILHAMENTO
VISIBILIDADE

INVISIBILIDADE

PUBLICAÇÃO DE DADOS

Um crescente número de novas modalidades de publicação está surgindo como resposta ao desafio de dar visibilidade e implementar estratégias de compartilhamento de dados de pesquisa. É importante observar que os mecanismos de publicação de dados tomam como solução um alinhamento ao sistema de reputação científica



As novas modalidades de publicação de dados e de suas representações descritivas demonstram com clareza que é possível de ancorar os sistemas de compartilhamento de dados às formas tradicionais de publicação, embora isso exija um alto grau de inovação e uma nova dinâmica que imponha mais velocidade nos processos de avaliação, que pode ser algo que se desenrole no tempo e se distribua no espaço de forma menos exclusiva (PAMPEL; DALLMEIR-TIESSEN, 2015).

A publicação dos dados de pesquisa como **objeto de informação independente**, em repositórios de dados ou centros de dados.

A publicação de **documentação textual** em ***data journal*** sobre dados de pesquisa na forma *de* ***data papers***

A publicação de dados de pesquisa **enriquecendo um artigo** por meio de *links* que podem ter valor semântico, nas chamadas **publicações ampliadas**

Publicação de dados de pesquisas de **experimentos que não deram certos e hipóteses não confirmadas** em periódicos voltados para essa condição

DATA paper journal

Uma publicação periódica científica cujo objetivo principal é **descrever coleções de dados** ao invés de reportar uma investigação científica

DESCREVE
os dados em forma legível por humanos

A **metodologia** sobre a qual os dados forma criados;

Detalha o **potencial de reuso** dos dados

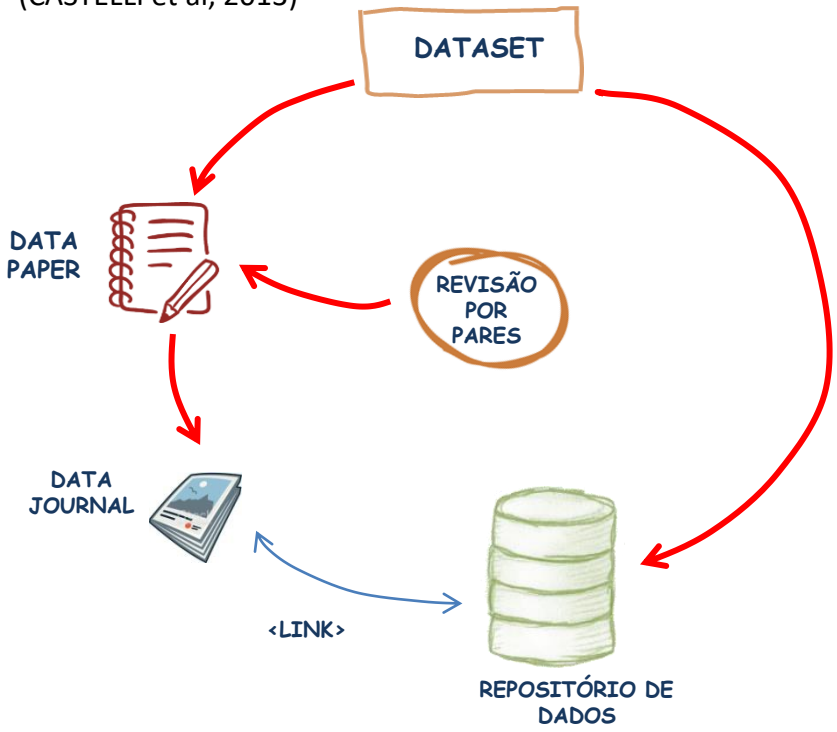
DESCREVE OS DADOS e não hipóteses ou argumentos desenvolvidos sobre os dados

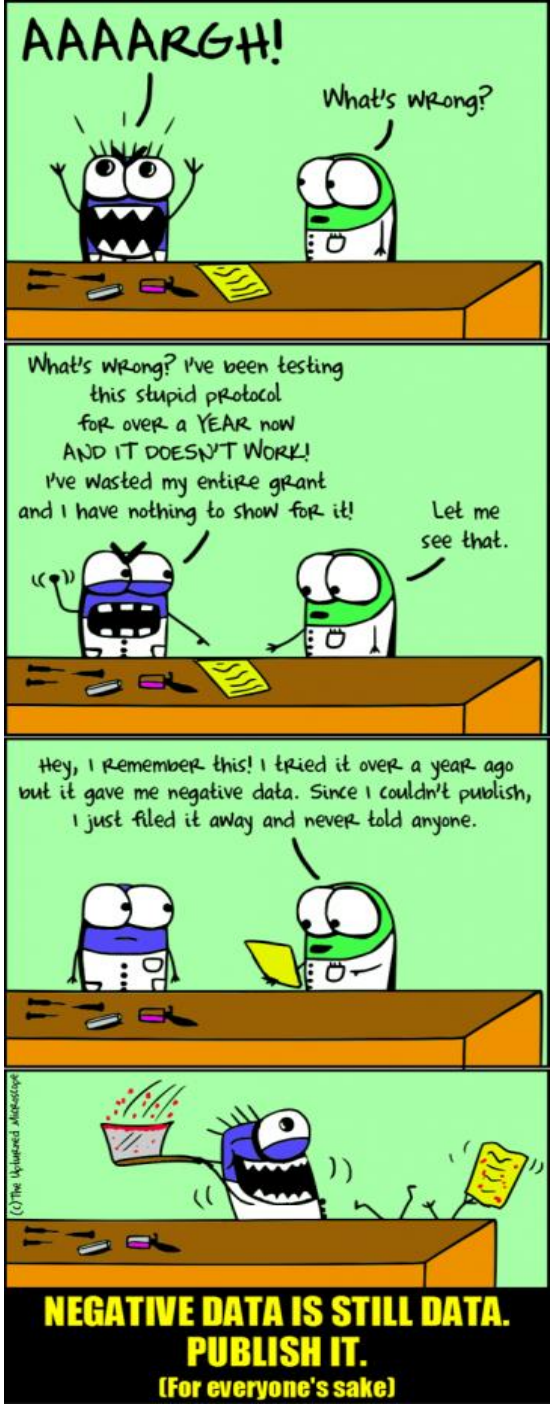
Oferecer uma publicação que **pode ser citada** e que dá **credito ao autor** e o outros envolvidos no processo;

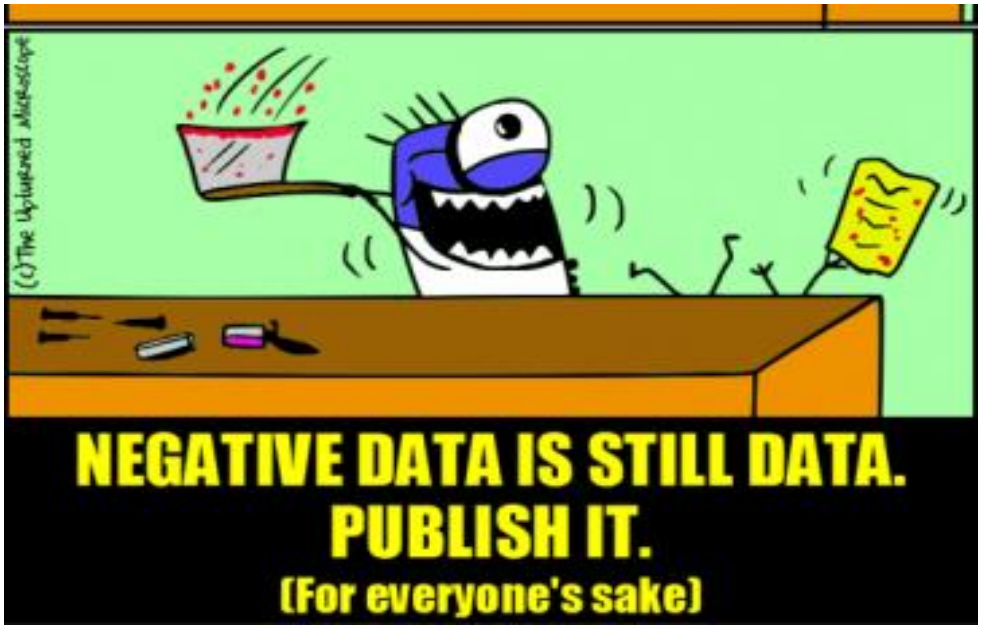
Assegura que os dados estejam **documentados para o reuso**;

Aumenta a visibilidade dos dados na comunidade científica

> A novidade interessante introduzida pelos *data journals* é que o modelo propõe um processo de publicação para dados que remete a publicação tradicional [...] A revisão por pares objetiva mensurar a originalidade e qualidade dos dados, ela é aplicada aos dados ao invés da publicação, e a sua "benção" é mandatória para os que os dados sejam publicados (CASTELLI et al, 2013)







O "viés de publicação do positivo" preocupa há décadas diversos pesquisadores. Partindo da ideia de que a comunidade científica só pode aprender com os resultados negativos se os dados forem publicados, existem alguns **periódicos científicos que investem na publicação do que não deu certo em diversas áreas**. Tais periódicos têm como premissa a concepção de que o suposto "fracasso" é tão importante na ciência como em outros aspectos da vida, e que o progresso científico não depende apenas das realizações de indivíduos isolados, mas requer colaboração, trabalho em equipe e comunicação aberta com todos os resultados, sejam eles positivos ou negativos.

Fonte: http://www.enago.com.br/blog/motivos-para-publicar-resultados-negativos/

JOURNAL OF NEGATIVE RESULTS
IN BIOMEDICINE

Journal
of Negative & No Positive Results

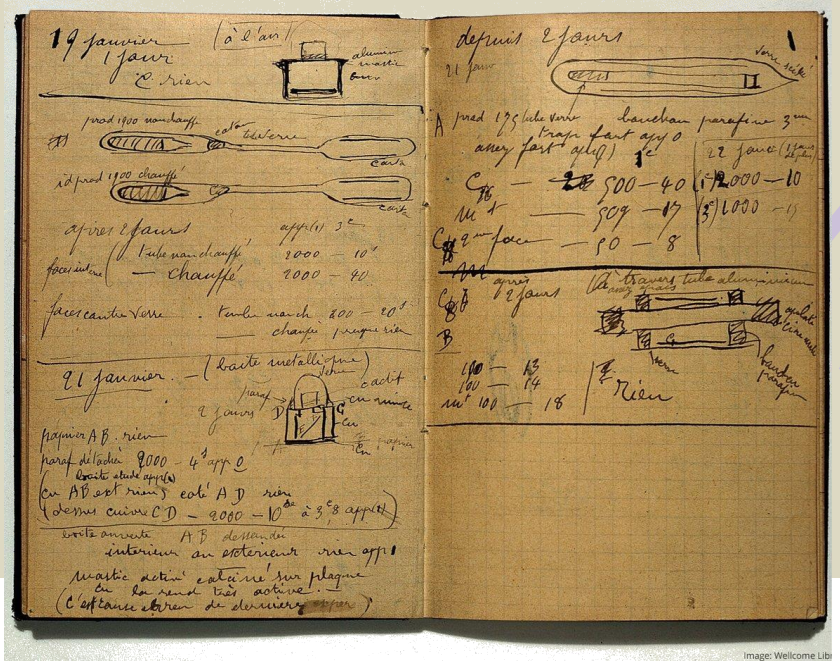
Journal of
Pharmaceutical
Negative Results

NEGATIVE RESULTS
SCIENTIFIC JOURNAL

PUBLICAÇÕES AMPLIADAS

Caderno de laboratório

O caderno de laboratório é uma ferramenta de organização e de memória que serve de registro primário da pesquisa científica e das atividades relacionadas. O caderno de pesquisa **registra as hipóteses, experimentos e análises iniciais ou interpretações dos experimentos;** serve também como o **registro legal da propriedade intelectual das ideias e dos resultados obtidos pela pesquisa** (SCHNELL, 2015).



Cadernos abertos
disponibilização dos dados acontece em tempo real, à medida que a pesquisa vai sendo feita

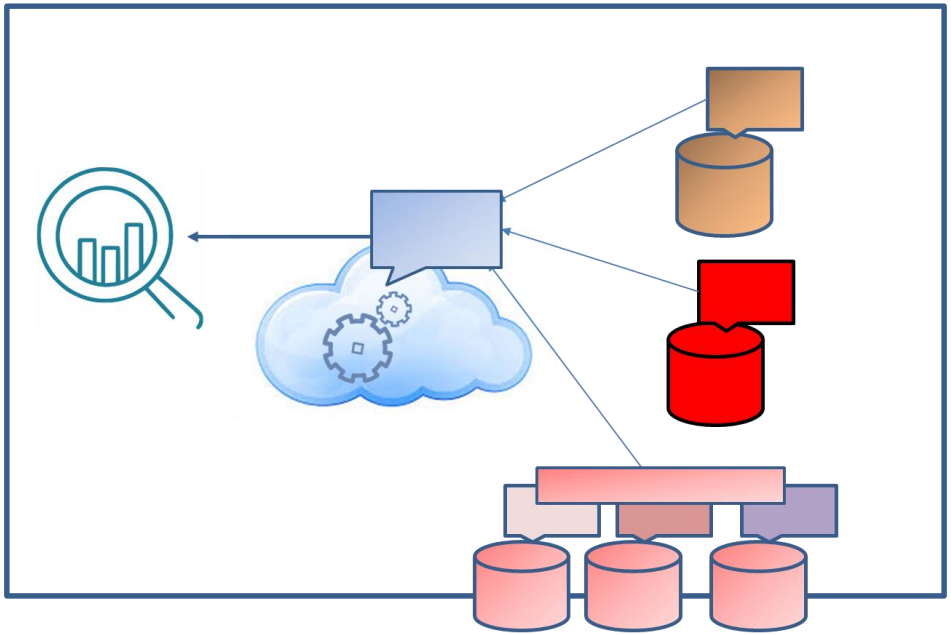
Sistemas complexos
integração com os equipamentos do lab

Cadernos Eletrônicos
auditoria | certificação

Cadernos convencionais

EXISTEM CÓDIGOS INTERNACIONAIS, NACIONAIS E INSTITUCIONAIS QUE DETALHAM AS ESPECIFICAÇÕES E GUARDA DESTES CADERNOS

PLATAFORMA GENÉRICA
ORIGEM TOP-DOWN
SEM CURADORIA
SEM PRESERVAÇÃO
SEM RECUPERAÇÃO
SEM SERVIÇOS
METADADOS GENÉRICOS



PLATAFORMA DISCIPLINAR
CULTURA DISCIPLINAR
SERVIÇOS ESPECÍFICOS
BUSCAS PRECISAS
METADADOS DISCIPLINARES
PROVENIÊNCIA
INSTRUMENTOS
NÍVEL DE PROCESSAMENTO
FLUXOS DA PESQUISA
METODOLOGIAS
INTERDISCIPLINARIDADE
ORIGEM COMUNITÁRIA

A GUISA DE CONCLUSÃO

O COMPARTILHAMENTO DE DADOS DA COMO PARTE DA CULTURA ACADÊMICA E A GESTÃO COMO PARTE DA PROFISSÃO DE PESQUISADOR

A BIBLIOTECA UNIVERSITÁRIA COMO PROTAGONISTA NA GESTÃO E INTEGRAÇÃO DE DADOS DA CAUDA LONGA

PRETEXTO PARA APROXIMAR A BIBLIOTECA DOS LABORATÓRIOS E DOS INFORMÁTICOS

NOVOS TEMAS DE PESQUISA