

# **R**epositório **D**igital de **D**ados de **P**esquisa

**Uma  
Experiência na  
Área de  
Ciências  
Nucleares**

**Luana Farias Sales – [lsales@ien.gov.br](mailto:lsales@ien.gov.br)**

Grupo de Pesquisa: Gestão do Conhecimento em Ciências Nucleares  
Comissão Nacional de Energia Nuclear –  
Instituto de Engenharia Nuclear  
Programa de Pós-graduação em Ciência da Informação PPGCI- IBICT

# Agenda

---

**Ciência orientada por dados:** nova configuração do fazer científico

**Dados de pesquisa:** insumos para inovação científica e tecnológica

**O que fazer com os dados de pesquisa?**

**Repositórios de dados de pesquisa:** vantagens e funções

**Carpe dIEN:** uma experiência na área de Ciências Nucleares

# Ciência orientada por dados

nova configuração do fazer científico

## eScience = uma nova forma de fazer ciência

Uma nova geração de **instrumentos científicos**, sensores, satélites, **softwares** de simulação, **colaboratórios** etc produzem em ritmo exponencial quantidades imensas e diversificadas de **dados de pesquisa**

“CIÊNCIA PRODUZIDA A PARTIR DO USO, ARMAZENAMENTO, PROCESSAMENTO, ANÁLISE E COMPARTILHAMENTO DE DADOS DE PESQUISA (GRAY, 2005).



**1ª Ciência experimental**

**2ª Ciência teórica**

**3ª Ciência baseada em simulação**

# Dados de pesquisa

insumos para inovação científica e tecnológica

## O QUARTO PARADIGMA CIENTÍFICO

# eScience

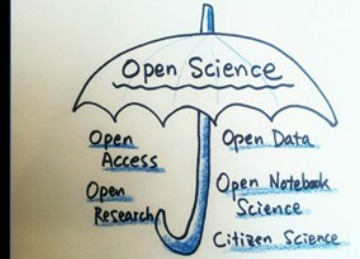
**ACELERA A PESQUISA CIENTÍFICA E GERA CONHECIMENTO COM BASE NA EXPLORAÇÃO DESSES DADOS**

Ferramentas de **software** e de **mineração** de dados ajudam a interpretar e transformar os dados brutos em **configurações ilimitadas** de informação e conhecimento.

# Ciência aberta

Movimento mundial que visa tornar a pesquisa científica disponível *on line* para todo o mundo.

“ Ciência aberta é a ideia de que o conhecimento científico de todos os tipos devem ser compartilhados abertamente tão logo o processo de descoberta seja iniciado. (NIELSEN, 2008)



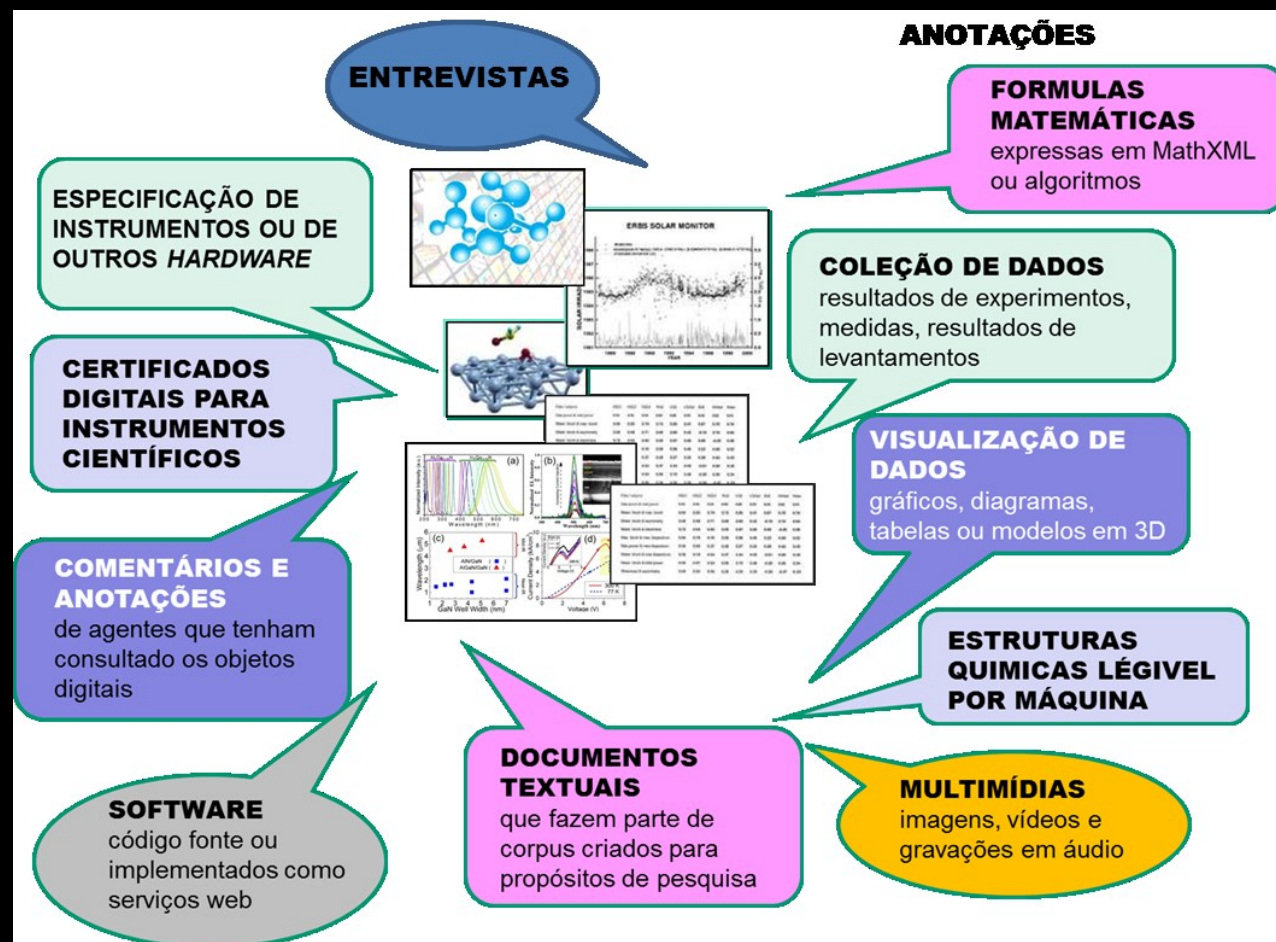
“ Se por um lado a eScience impulsiona a geração de dados, por outro a Ciência aberta amplia o alcance e a disseminação dos dados.



# DILÚVIO DE DADOS DE PESQUISA



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



**POLÍTICAS DE AMPLO ESPECTRO**

**OPÇÕES GERENCIAIS E TECNOLÓGICAS PARA O ARQUIVAMENTO CURADORIA DIGITAL**

# ORIGENS DOS DADOS

## DADOS OBSERVACIONAIS

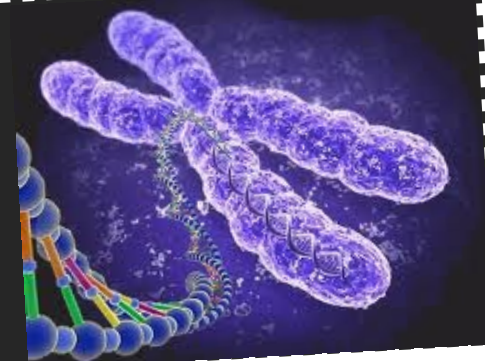
são provenientes de situações não controladas, como erupção de um vulcão numa data específica, a atitude dos eleitores ou fotografia de uma supernova – que constituem enfim registros históricos que não podem ser coletados uma segunda vez e, portanto, são armazenados e arquivados para sempre

**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





## **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.



# DADOS DE PESQUISA SÃO MUITO SUSCETÍVEIS A PERDAS

# PUBLICAÇÃO

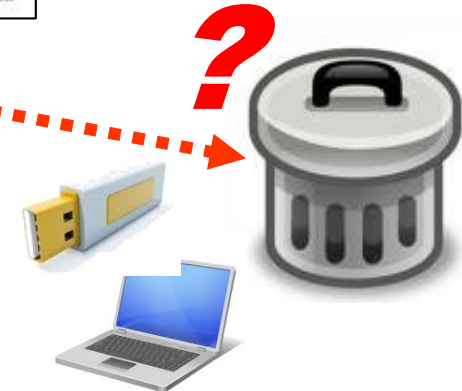
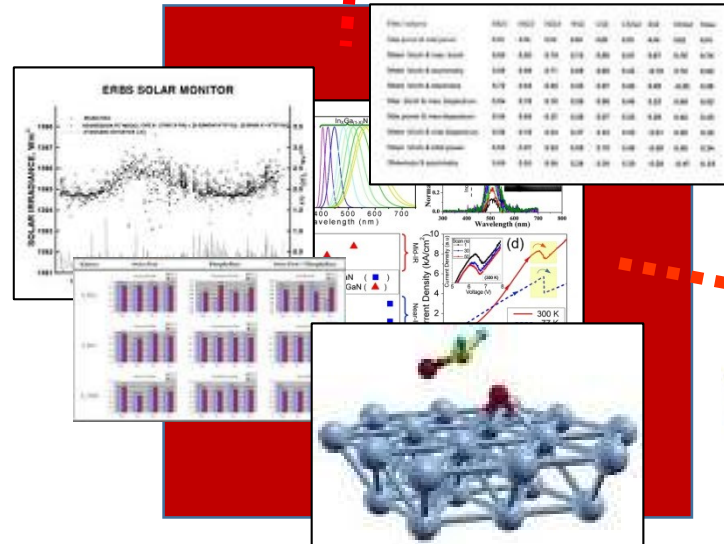
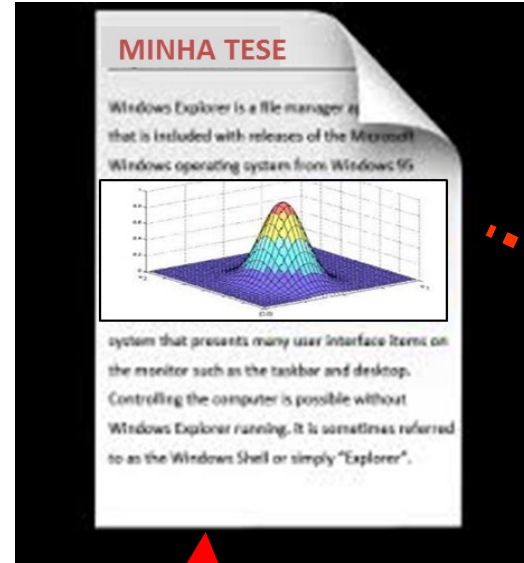
# PESQUISA TESE

# DADOS DIGITAIS

BIBLIOTECA  
CONVENCIONAL

**BIBLIOTECA DE TESES  
E DISSERTAÇÕES**

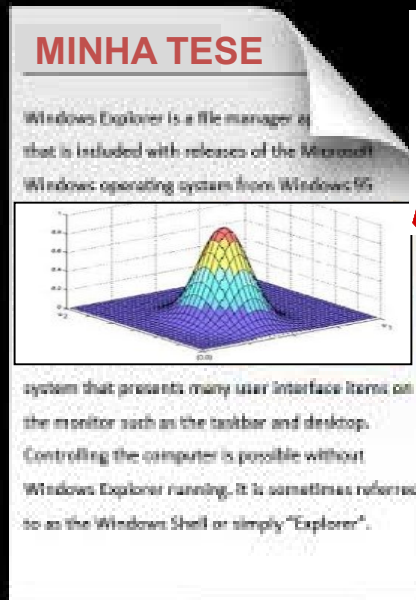
## REPOSITÓRIOS





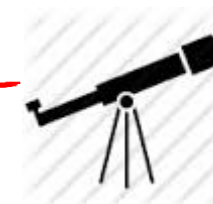
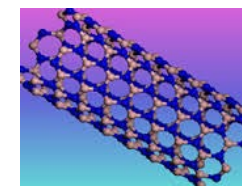
# VISÍVEL INVISÍVEL

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

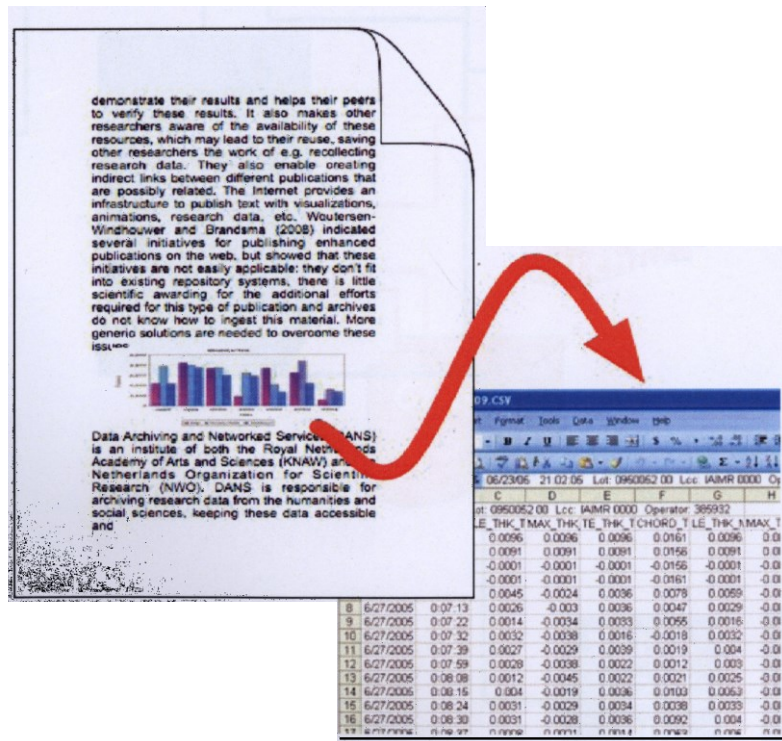


## UMA VISÃO DOS DADOS

**[revisão por pares]**  
**[validação da pesquisa]**

[illegible]

# REUSO DOS DADOS DE PESQUISA



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

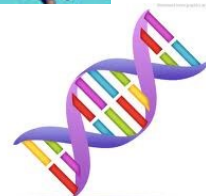
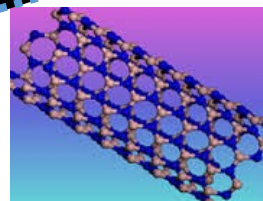
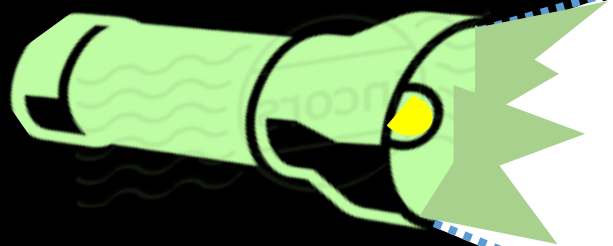
pesquisadores em bioinformática combinam coleções de dados originados no domínio da biologia, genética e engenharia

**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**



“

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.

# O que fazer com os dados de pesquisa?

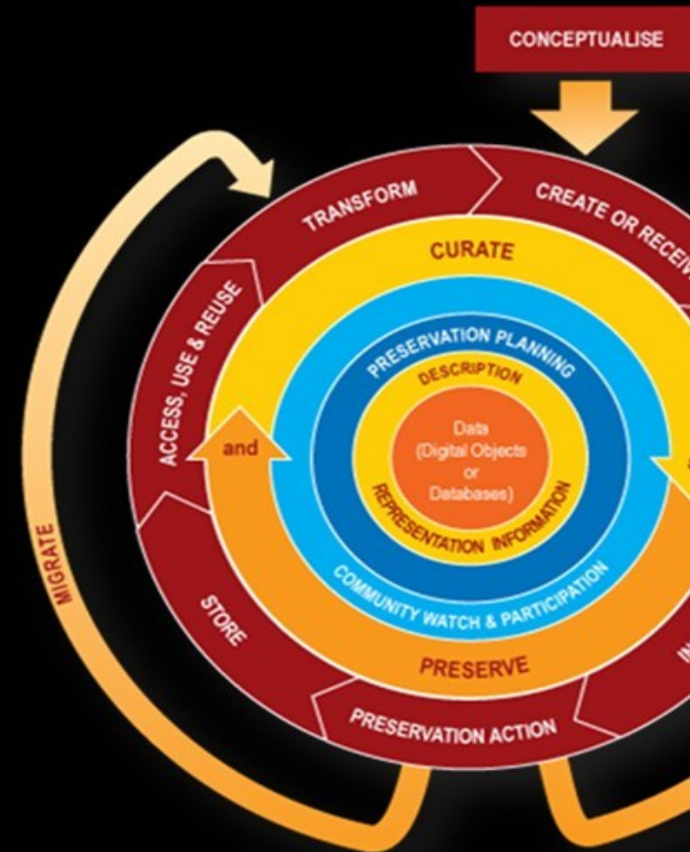
# CURADORIA DIGITAL DE DADOS DE PESQUISA

“

Manutenção, preservação e agregação de valor a dados de pesquisa **durante o seu ciclo de vida.** (DCC, 2003)

“

Todas as atividades envolvidas na **gestão de dados**, desde o planejamento de sua criação – quando os sistemas são projetados – passando pelas boas práticas de digitalização, na seleção dos formatos e da documentação, e na garantia de estarem sempre adequados para serem **descobertos e reusados agora e no futuro**. (ABBOTT, 2008)





# REPOSITÓRIOS DE DADOS DE PESQUISA

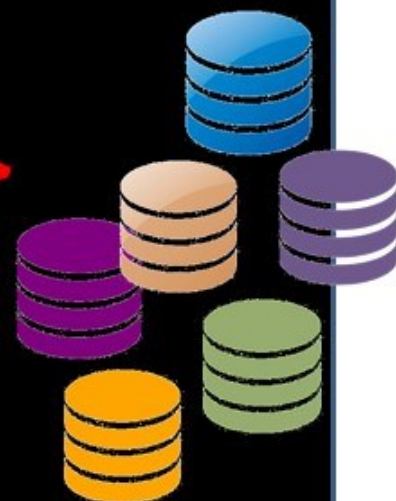
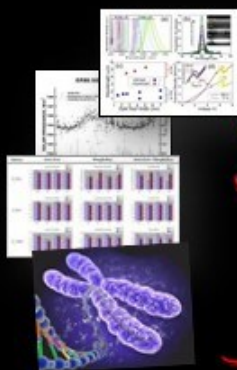


DADOS DE DADOS CIENTÍFICOS DIGITAIS QUE  
GARANTEM O ACESSO A RESULTADOS DE PESQUISA  
AGORA E NO FUTURO, TENDO COMO PERSPECTIVA  
PRIMORDIAL O ACESSO ABERTO AO QUE É  
CONSIDERADO TAMBÉM UM PATRIMÔNIO DIGITAL DA  
HUMANIDADE  
(SURF FOUNDATION, 2013)



# VANTAGENS

DADOS



REPOSITÓRIOS  
DE DADOS

GARANTIA DE QUE OS DADOS ESTÃO DE ACORDO COM **PADRÕES DE QUALIDADE**

**PRESERVAÇÃO DE LONGO PRAZO**

**ARMAZENAMENTO SEGURO** COM CAPACIDADE DE **CONTROLAR O ACESSO**, SEMPRE QUE NECESSÁRIO

**BACK-UPS** DE DADOS REGULARES

**DESCOBERTA DE RECURSOS** ON-LINE ATRAVÉS DE CATÁLOGOS DADOS

ACESSO AOS DADOS EM **FORMATOS POPULARES**

**ACORDOS DE LICENCIAMENTO** PARA RECONHECER OS DIREITOS ASSOCIADOS AOS DADOS

MECANISMO DE **CITAÇÃO PADRONIZADA**

**VISIBILIDADE DOS DADOS** PARA MUITOS USUÁRIOS

- MONITORAMENTO DO USO SECUNDÁRIOS DOS DADOS

OS CENTROS DE DADOS, COMO QUALQUER ARQUIVO TRADICIONAL, GERALMENTE SE APLICAM ALGUNS CRITÉRIOS PARA AVALIAR E SELECIONAR DADOS PARA PRESERVAÇÃO.



# SABORES



REPOSITÓRIO INSTITUCIONAL DE DADOS DE PESQUISA



REPOSITÓRIO TEMÁTICO DE DADOS DE PESQUISA



REPOSITÓRIO MULTIDISCIPLINAR DE DADOS DE PESQUISA



REPOSITÓRIO DE DADOS DE PESQUISA DE UM PROJETO ESPECÍFICO

# Gestão de Dados de Pesquisa

## fases



- **CAPTURA DE DADOS**

Coleta e Seleção de coleções de dados – primários ou derivados - provenientes de experimentos, simulações, observações, questionários, levantamentos etc. Os dados podem ser submetidos pelos próprios autores ou por equipes especializadas vinculadas ao serviço.

- **CATALOGAÇÃO DAS COLEÇÕES DE DADOS**

Descrição, atribuição de metadados e inclusão de documentação que assegurem que os dados possam ser acessados e interpretados no tempo e no espaço.

- **ARQUIVAMENTO E PRESERVAÇÃO -**

Arquivamento seguro que garante a gestão de curto e longo prazo das coleções de dados orientadas por um plano/política de preservação digital

- **INTEROPERABILIDADE**

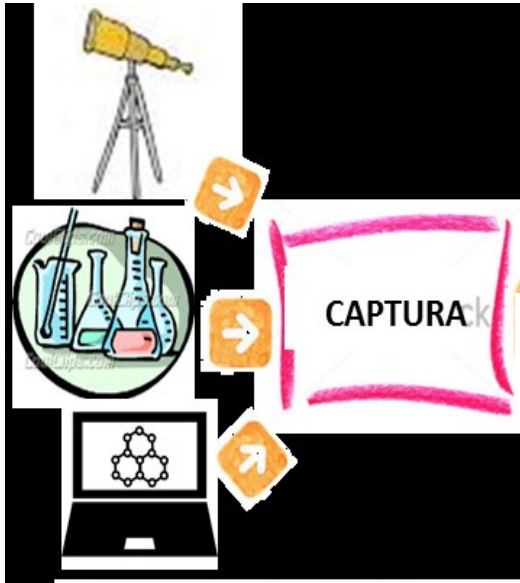
Intercâmbio e compartilhamento e *linkage* com outros repositórios de dados e outros sistemas de informação (repositórios institucionais, bibliotecas digitais de publicações acadêmicas, editoras científicas)

- **RECUPERAÇÃO, ACESSO E REUSO**

Interface web para a descoberta, acesso e *download* de coleções de dados relevantes para o usuário ou para aplicações computacionais, como visualização e mapeamento, que podem prover serviços a partir dessas coleções; vinculado a uma política de acesso estabelecida pela instituição que inclui: tempo de embargo, direito de acesso, pagamentos, restrições sobre determinadas coleções, acesso somente aos metadados, registros de usuários e termos de uso dos dados.

# FUNÇÕES DO REPOSITÓRIO DE DADOS

## CAPTURA DE DADOS



### Desenvolvimento de coleções de dados (Data appraisal)

- Coleta e Seleção dos dados
- Verificação do enquadramento no escopo do repositório;
- Verificação dos formatos de arquivos aceitáveis para submissão;
- Verificação dos direitos associados às coleções (copyright e licenças);
- Verificação de dados sensíveis (dados não anonimizados, confidenciais, pessoais);
- Verificação do volume e quantidade de arquivos;
- Verificação dos metadados gerais e disciplinares que acompanham os dados;
- Normalização para elenco de formatos padronizados aceitos para arquivamento e disseminação;
- Controle de qualidade dos dados;
- Definição de tempo de embargo.

# FUNÇÕES DO REPOSITÓRIO DE DADOS

## CATALOGAÇÃO DAS COLEÇÕES DE DADOS



- Atribuição de metadados descritivos, estruturais, administrativos, técnicos (que inclui os relativos às dependências técnicas dos objetos digitais)
- Atribuição de metadados de preservação, que assegurem a proveniência, autenticidade e integridade dos dados ao longo do tempo;
- Uso de taxonomias especializadas e disciplinares;
- Atribuição de identificador persistente (DOI, Handles, UNF, URN, etc.) que permita que os dados possam ser localizados de forma persistente e citados como as publicações acadêmicas;
- Identificação do autor (ORCID ID, Scopus Author ID, ResearcherID etc.)
- Inclusão de documentação sobre os dados, incluindo descrição do projeto, dos arquivos e dos parâmetros; cadernos de laboratório e de campo, protocolos de pesquisa ou metodologia etc.;
- Vinculação (por links) a publicações e a dados relacionados internos e externos ao repositório



# FUNÇÕES DO REPOSITÓRIO DE DADOS

## ARQUIVAMENTO E PRESERVAÇÃO

---



- Armazenamento em sistemas seguros;
- Gestão da preservação de curto prazo (backups, backups redundantes offsite; checagem de integridade, armazenamento seguro, criptografia, compressão);
- Gestão de longo prazo (migração, emulação, reformatação para formatos padronizados, aplicação de normas pertinentes (OAIS, TRAC), informação de fixidade voltada para validar a autenticidade e integridade de um objeto digital (checksums, assinatura digital);
- Implementação de trilhas de auditoria;
- e compartilhamento e *linkage* com outros repositórios de dados e outros sistemas de informação (repositórios institucionais, bibliotecas digitais de publicações acadêmicas, editoras científicas)

# FUNÇÕES DO REPOSITÓRIO DE DADOS

---



## INTEROPERABILIDADE

- Intercâmbio
- Disponibilização de metadados segundo o protocolo OAI-PMH;
- Agregação para formação de publicações ampliadas segundo o padrão OAI-ORE;
- Uso dos padrões, *web service*, *linked data* e outros
- Empacotamento de metadados para intercâmbio segundo o padrão METS.



## RECUPERAÇÃO, ACESSO E REUSO

- Disponibilização de interfaces web para recuperação, acesso e *download*;
- Oferta de aplicações e serviços sobre as coleções;

# Re3data.org

## Registry of Research Data Repositories Initiative

- Diretório que reúne a **descrição de repositórios digitais de dados** de pesquisa em uma grande base de dados.
- Objetivo: **descrever, indexar e estruturar repositórios de dados** de todos os domínios em um único registro web.
- Oferece **orientação aos pesquisadores** no seu papel de produtores de dados bem como na qualidade de usuários de dados.

	ICONE	SIGNIFICADO
Informação		O repositório disponibiliza informações adicionais sobre os seus serviços
Acesso		O repositório oferece acesso aberto aos seus dados
		O repositório oferece acesso restrito aos seus dados
		O repositório oferece acesso fechado aos seus dados
Licenças		Os termos de uso e licenças dos dados são disponibilizados pelo repositório
Identificador Persistente		O repositório usa DOI para tornar seus dados persistentes, únicos e citáveis
		O repositório usa URN para tornar seus dados persistentes, únicos e citáveis
		O repositório usa ARK para tornar seus dados persistentes, únicos e citáveis
		O repositório usa HANDLES para tornar seus dados persistentes, únicos e citáveis
		O repositório usa PURL para tornar seus dados persistentes, únicos e citáveis
		O repositório usa outros esquemas de identificação para tornar seus dados persistentes, únicos e citáveis
Certificados e Padrões		O repositório é certificado ou segue os padrões para repositórios
Política		O repositório possui e disponibiliza um documento de política

# Carpe dIEN

como surgiu?



**JUSTIFICATIVAS:** Necessidade de **mecanismos de transparência** que testemunhem as **atividades** que se desenrolam nos **laboratórios**, tendo em vista que investimentos volumosos são feitos por parte da sociedade, que agora exige retornos visíveis.

Necessidade de desenvolvimento de **espaços virtuais de memória** onde dados e informações digitais possam ser tratados, armazenados e preservados para acesso de longo prazo por parte de todas as comunidades interessadas.

Necessidade de criação de uma “**ciberinfraestrutura**”, que integrasse **várias tecnologias** e metodologias, que melhore a qualidade da pesquisa e torne mais veloz a **transformação desse conhecimento em benefícios para a sociedade**.

Necessidade de **ferramentas** e espaços virtuais que intensifiquem a **interatividade, o compartilhamento de recursos e o diálogo entre pares**, encurtando o ciclo da comunicação científica e ampliando a visibilidade e o alcance das pesquisas científicas.

**OBJETIVO GERAL:** Desenvolver uma infraestrutura informacional, gerencial e tecnológica voltado para a Gestão do Conhecimento e preservação da memória científica digital do Instituto de Engenharia Nuclear

## OBJETIVOS ESPECÍFICOS:

Promover a preservação da memória científica digital do IEN.

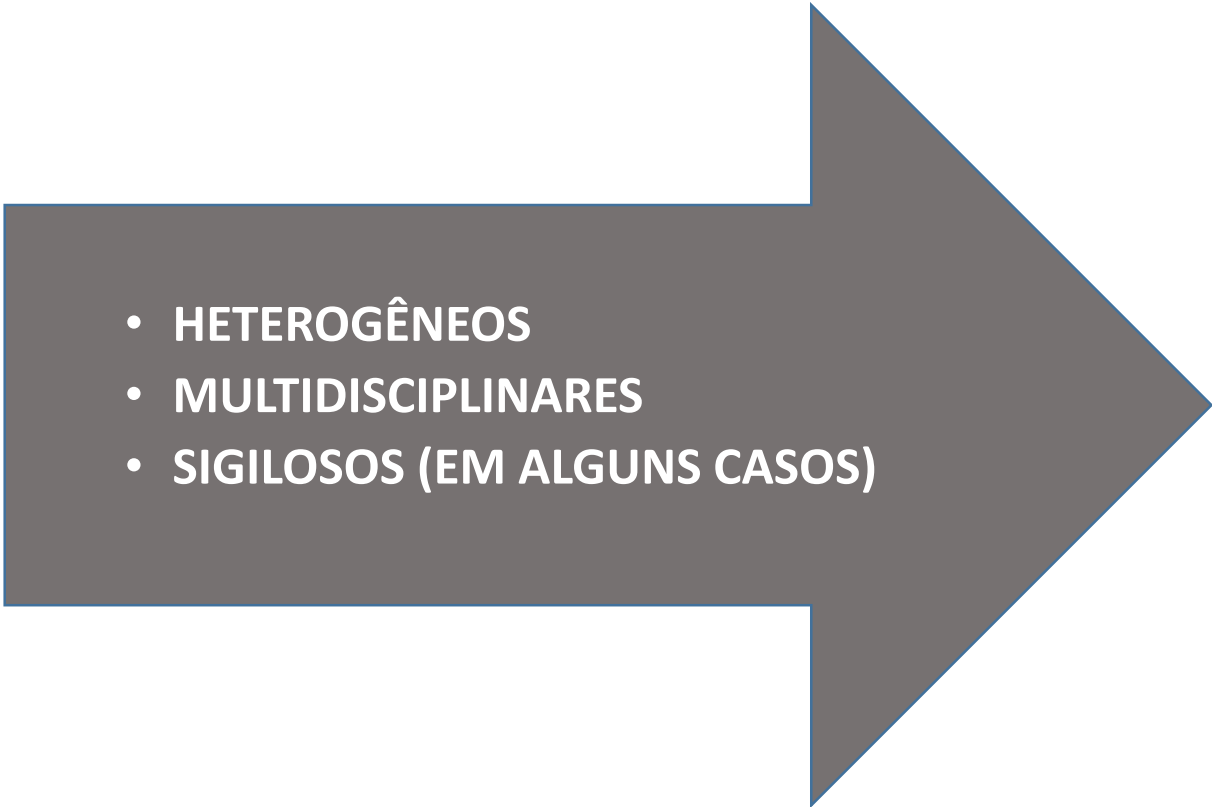
Servir de instrumento para a implementação de uma política de gestão de CT&I.

Valorizar e disseminar a produção do corpo técnico da Instituição.



# Como são os dados no IEN?

---

- 
- **HETEROGÊNEOS**
  - **MULTIDISCIPLINARES**
  - **SIGILOSOS (EM ALGUNS CASOS)**

## **POR ORIGEM DOS DADOS:**

Observacionais  
Computacionais  
Experimentais  
Outros

## **ESTÁGIO DE GERAÇÃO DOS DADOS**

Bruto  
Semi-processado  
Processado  
Derivado  
Outros

## **NATUREZA DOS DADOS**

Multimídia  
Numérico  
Textual  
Software  
Visualização  
Artefato  
Processo  
Simulação

# Carpe dIEN

Como está organizado?

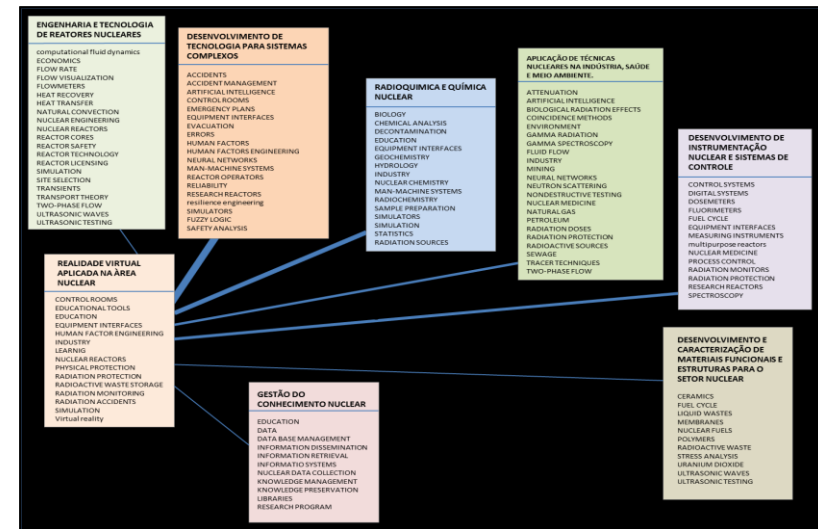
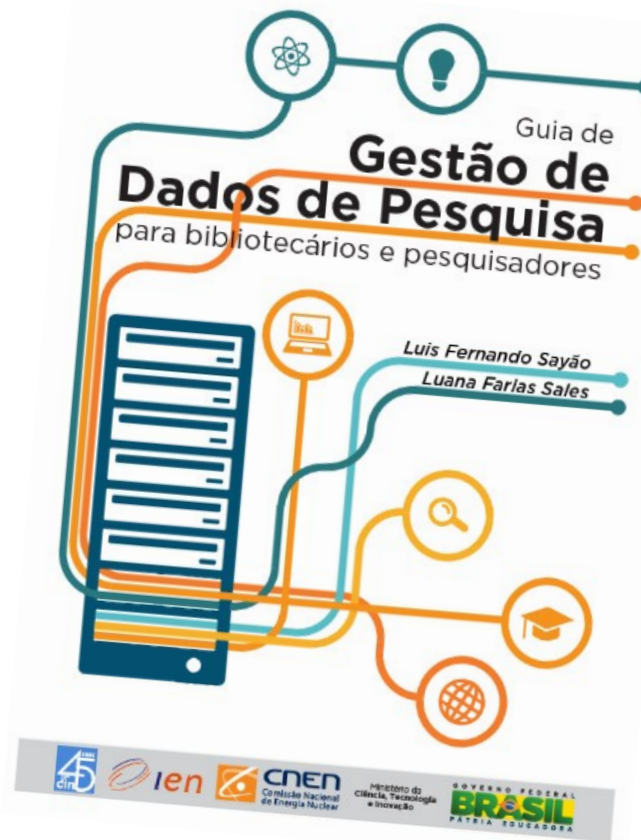
- Tecnologia adotada: Dspace
- Está dividido por áreas temáticas de pesquisa do Instituto
- Cada área está subdividida em coleções, na qual uma é a de “Dados de pesquisa”



# Carpe diem

## o que já foi feito?

- Modelagem genérica do repositório para recebimento de dados
- Divulgação institucional
- Mapeamento dos tipos de dados gerados pelos projetos de pesquisa em andamento
- Criação de política para o repositório
- Elaboração de Manual genérico para o repositório
- Desenvolvimento de coleções – em andamento
- Criação de Guia para Gestão de Dados de Pesquisa
- Inserção, teste e remodelagem de cada coleção do repositório destinada a receber dados



### REPOSITÓRIO DIGITAL DO IEN METADADOS PARA DADOS DE PESQUISA – ELEMENTOS DE PREENCHIMENTO – VERSÃO 2.0 – 13/03/12 / Atualizado em 17 de março 2013

TÍTULO (DC: title)							Definição/Usu/Exemplo	
Elemento DC	PhI ID	Dspace ID	Qualificador DC	Etiqueta	Rep ?	Obr ?		
Title	018 181	64		TÍTULO	não	sim	Definição: Nome atribuído ao recurso Uso: Entre com o texto original, ordem e a ortografia do título do recurso. Exemplos: Para substituir, separe-o do título por dois pontos. Para substituir, separe-o do título por dois pontos. Exemplos: Nem cometas nem tãmboreis: um estudo	Definição: Nome atribuído ao recurso Uso: Entre com o texto original, ordem e a ortografia do título do recurso. Exemplos: Para substituir, separe-o do título por dois pontos. Para substituir, separe-o do título por dois pontos. Exemplos: Nem cometas nem tãmboreis: um estudo
	019 65		alternativo	TÍTULO ALTERNATIVO	não	não		Definição: Formas alternativas que aparecem no título, como abreviaturas, traduções, aumento do título, etc.

AUTOR (DC: creator)							Definição/Usu/Exemplo	
Elemento DC	PhI ID	Dspace ID	Qualificador DC	Etiqueta	Rep ?	Obr ?		
Creator	016 017	9		AUTOR	sim	sim	Definição: Principal entidade responsável pela criação do recurso. Uso: Utilizar a forma invertida do nome para autores pessoais. Para autor coletivo use a forma mais conhecida na forma direta Exemplo: Sales, Luana Farias; Instituto de Engenharia Nuclear	Definição: Principal entidade responsável pela criação do recurso. Uso: Utilizar a forma invertida do nome para autores pessoais. Para autor coletivo use a forma mais conhecida na forma direta Exemplo: Sales, Luana Farias; Instituto de Engenharia Nuclear
			departament	DEPARTAMENTO	sim	não	Definição: Departamento ou similar de vinculação do autor	Definição: Departamento ou similar de vinculação do autor
			researchgroup	GRUPO DE PESQUISA	sim	não	Definição: Grupos de pesquisa de vinculação do autor	Definição: Grupos de pesquisa de vinculação do autor
			e-mail	E-MAIL	sim	não	Definição: E-mail do autor	Definição: E-mail do autor
	72		curriculum	LATRES	não	não	Definição: Link para o Currículo Lattes do autor	Definição: Link para o Currículo Lattes do autor
			nameidentifier	IDENTIFICADOR DE NOMES	não	não	Definição: Identificador único de nomes	Definição: Identificador único de nomes
			namesidentifier	ESQUEMA DE IDENTIFICADOR NOMES	não	não	Definição: nome e/ou URL do esquema de identificador de nomes	Definição: nome e/ou URL do esquema de identificador de nomes

# Carpe diem

---

## desafios

- Convencimento dos pesquisadores sobre a importância do depósito de seus dados
- Ausência de uma política institucional tornando o depósito de dados compulsório
- Necessidade de uma infraestrutura para DOI.
- Ausência de espaço para armazenamento de grandes quantidades de dados
- Ausência de um vocabulário controlado para padronizar a indexação temática dos dados e permitir interoperação semântica
- Ausência de pessoal capacitado para lidar com esse tipo de atividade e manutenção dos já capacitados
- Grande proliferação de metadados, tendo em vista a necessidade de criar novos a cada conjunto de dados diferentes que precisam ser tratados



# Carpe diEM

---

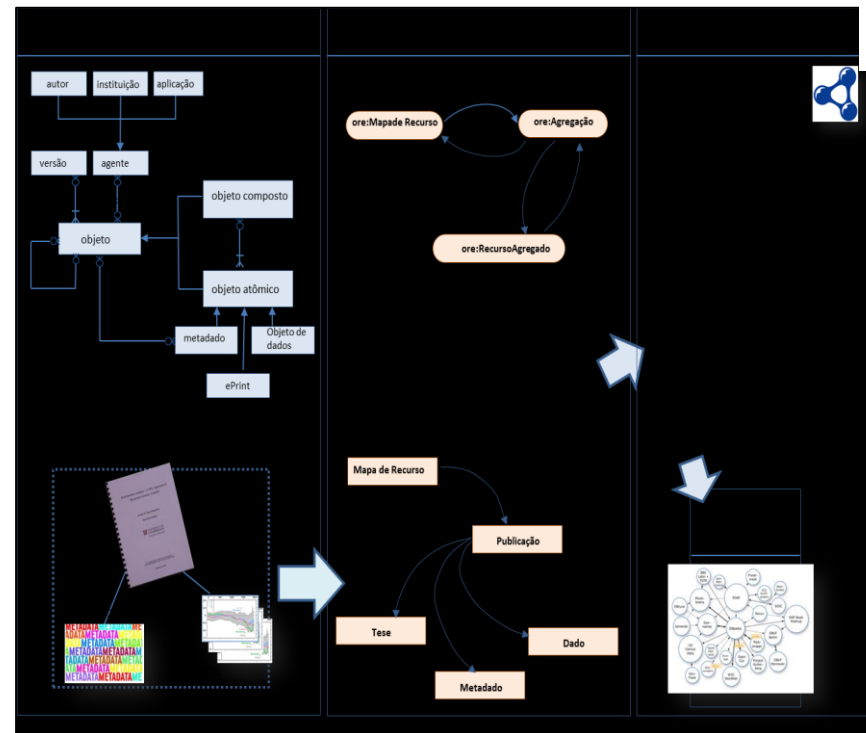
## estágio atual

- Levantamento dos dados de cada projeto
- Inserção de, pelo menos, uma amostra de cada tipo de dado identificado nos projetos
- Criação de um manual de inserção de dados para ser disponibilizado na página principal do repositório
- Cadastro no Re3data.org
- Modelagem e teste de novas plataformas de gestão de dados de pesquisa (Dataverse e Ckan)

# Publicação Ampliada

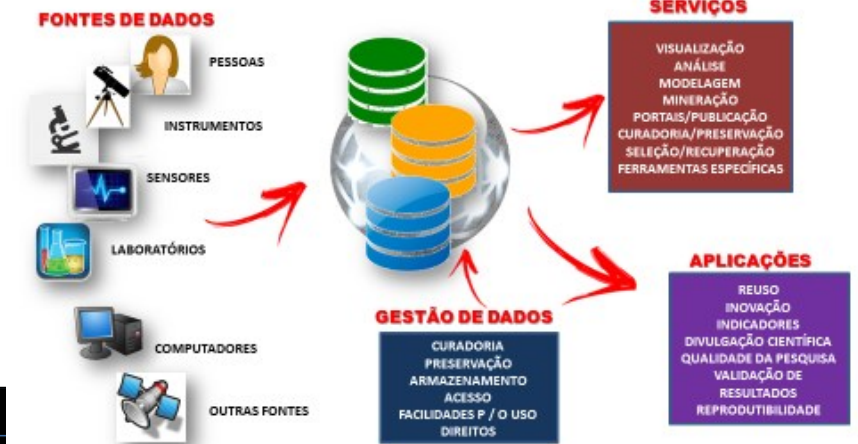
## um sonho possível...

- Integrar em uma interface única publicações, dados de pesquisa e outras informações úteis para os pesquisadores reconstituindo a memória de pesquisa dos projetos desenvolvidos no IEN
- Criar serviços inovadores sob a plataforma, como um espaço de colaboração entre os pesquisadores
- Expandir o modelo para toda a área de Ciência Nucleares



## CIBERINFRAESTRUTURA PARA DADOS DE PESQUISA

CONSISTE DE SISTEMAS DE COMPUTAÇÃO, SISTEMAS DE ARMAZENAMENTO, INSTRUMENTOS AVANÇADOS E REPOSITÓRIOS DE DADOS, AMBIENTES DE VISUALIZAÇÃO E PESSOAS, TUDO LINKADO POR REDES DE ALTO DESEMPENHO QUE TORNA POSSÍVEL **INOVAÇÃO CIENTÍFICA E DESCOBERTAS** QUE DE OUTRA MANEIRA NÃO SERIA POSSÍVEL



# À GUIA DE CONCLUSÃO

---

- Existem outros tipos de informações que merecem ser preservados e disseminados via tecnologia de repositórios
- No ambiente científico, dados de pesquisas são insumos importantes para a geração de novos conhecimentos
- Apenas a implementação de repositórios não vai solucionar o problema, é preciso muita gestão, curadoria e toda uma infraestrutura que permita que publicações e dados se interliguem
- Mas o repositório pode ser um começo...