PRINCÍPIOS FAIR APLICADOS À REPOSITÓRIOS

Luana Farias Sales
IBICT – PPGCI
GOFAIR - Brasil
luanafsales@gmail.com

3 PERGUNTAS E 1 DESAFIO

- Afinal, o que são princípios FAIR?
- Como tornar seus dados FAIR?
- Por que aplicar princípios FAIR à repositórios?
- Dos repositórios à plataformas de gestão de dados e serviços FAIR

A SAGA DOS DISPOSITIVOS DE MEMÓRIA ATÉ O DIGITAL (DE 1930 AOS DIAS DE HOJE): O DIÁLOGO POSSIVEL

1930

Que tal um **REPERTÓRIO** que reúna toda memória humana e a torne acessível a qualquer estudante em qualquer parte do mundo?

WELLS

Ótima idéia! Podemos reunir pedaços de informações em fichas 3x5 e criar um motor de busca analógico! Ou ainda criar uma cidade do conhecimento, que chamaremos **MUNDANEUM!!!**

OTLET

1945

E que tal um dispositivo onde os indivíduos possam armazenar todos os seus livros, dados e informações?. O nome será **MEMEX!!!**

BUSH

1960

Ótimo, Bush! E quem sabe interligarmos essas informações através de elos (links). Podemos chamar isso de **HIPERTEXTO!!!**

TED NELSON

1960

Ted, podemos ainda acrescentar a idéia de homens e computadores cooperarem na tomada de decisões e no controle de situações complexas. Vamos criar a **ARPANET!!!**

LICKLIDER

1980

Bom Pessoal, tendo computadores, podemos **DISPENSAR O PAPEL**, certo?

LANCASTER

Com tantas tecnologias disponíveis podemos construir uma Internet de Dados & Serviços FAIR

BAREND MONS

1950

Para tudo isso funcionar temos que ter métodos eficazes de **RECUPERAÇÃO DE INFORMAÇÃO**

MOOERS

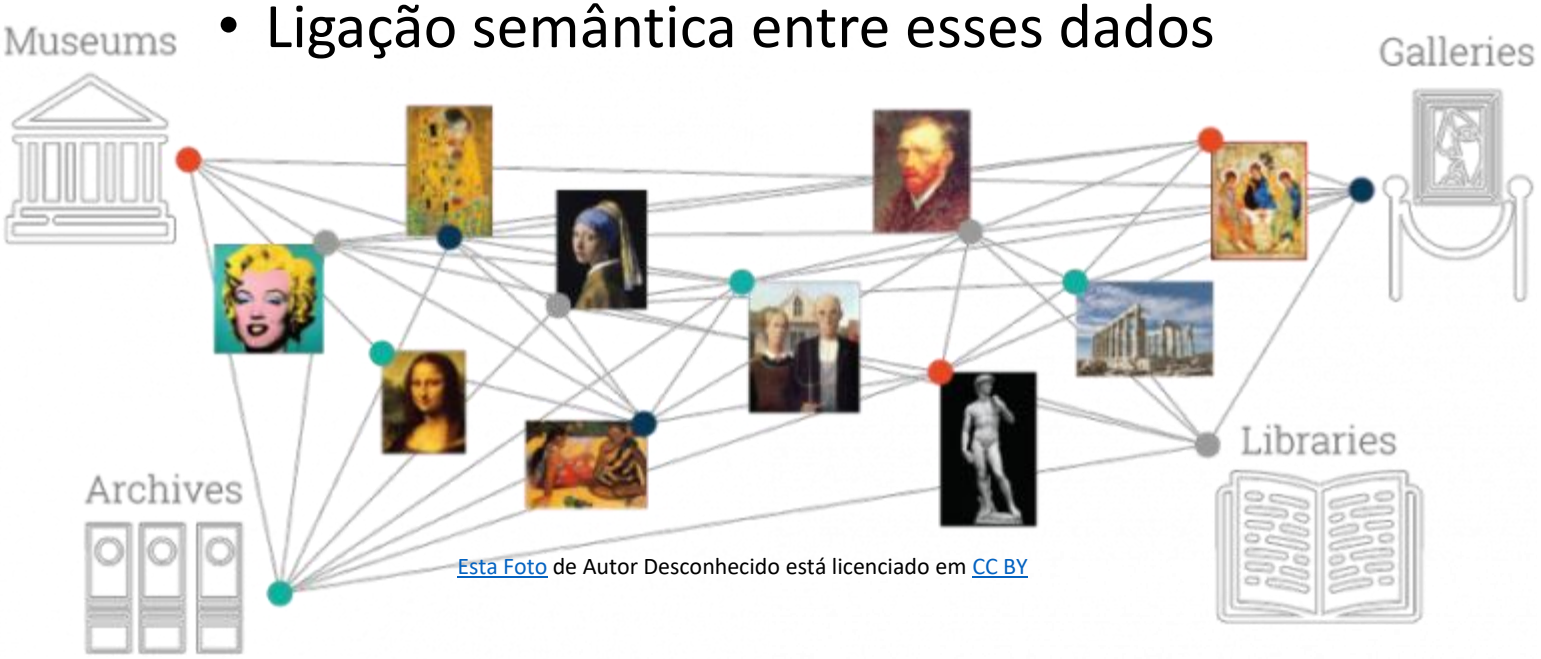
BARNES-LEE

Sim!!! Depois podemos criar uma grande teia de informações, a **WORLD WIDE WEB**, que poderá ser visualizada por **BROWSERS**!!!

Podemos então sonhar com uma **WEB SEMÂNTICA** em que máquinas consigam processar conhecimento!!!!

A Web Semântica = Uma Web de Dados

- Um sonho comum
- Uma realidade que pode acontecer a partir de:
  - disponibilização de (meta)dados de qualidade
  - Ligação semântica entre esses dados



AVALIAÇÃO 5 ESTRELAS DE BARNES-LEE

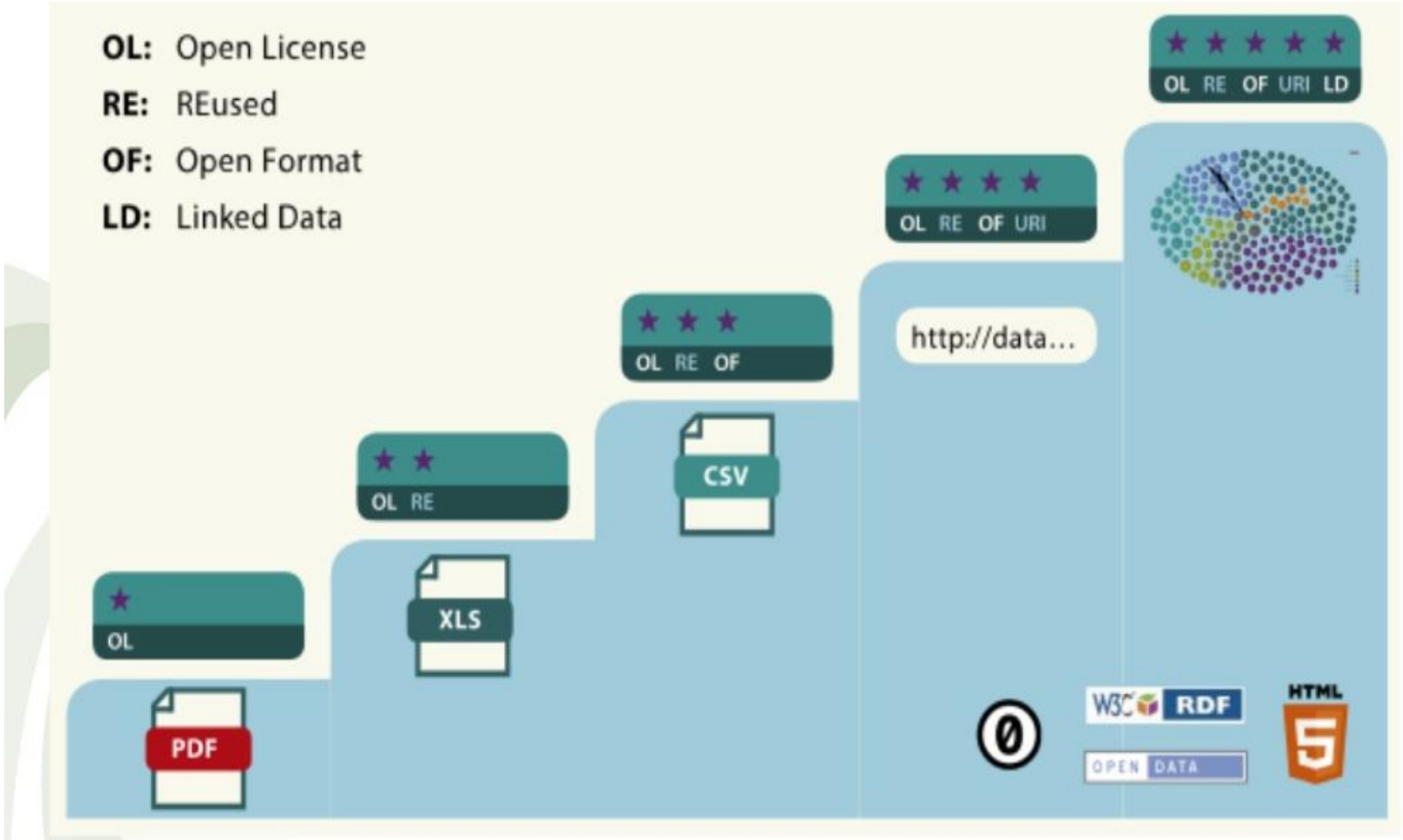
1ª Estrela - atribuída aos dados que são publicados sob **licença aberta** (Open License - OL), independente do formato em que se apresenta;

2ª Estrela - atribuída aos dados que além de publicados sob licença aberta são **estruturados e legíveis por máquinas** (Readable Machine - RE);

3ª Estrela - atribuída aos dados que são publicados em **formato aberto** não proprietário (Open Format - OF), sendo possível a manipulação dos dados sem a necessidade de uso de um software proprietário;

4ª Estrela - atribuída aos dados que possuem as classificações anteriores e que utilizam **Identificadores Uniforme de Recursos** (Uniform Resource Identifier - URI) para nomear os dados, permitindo criar ligações que façam reuso dos dados disponibilizados na web; e

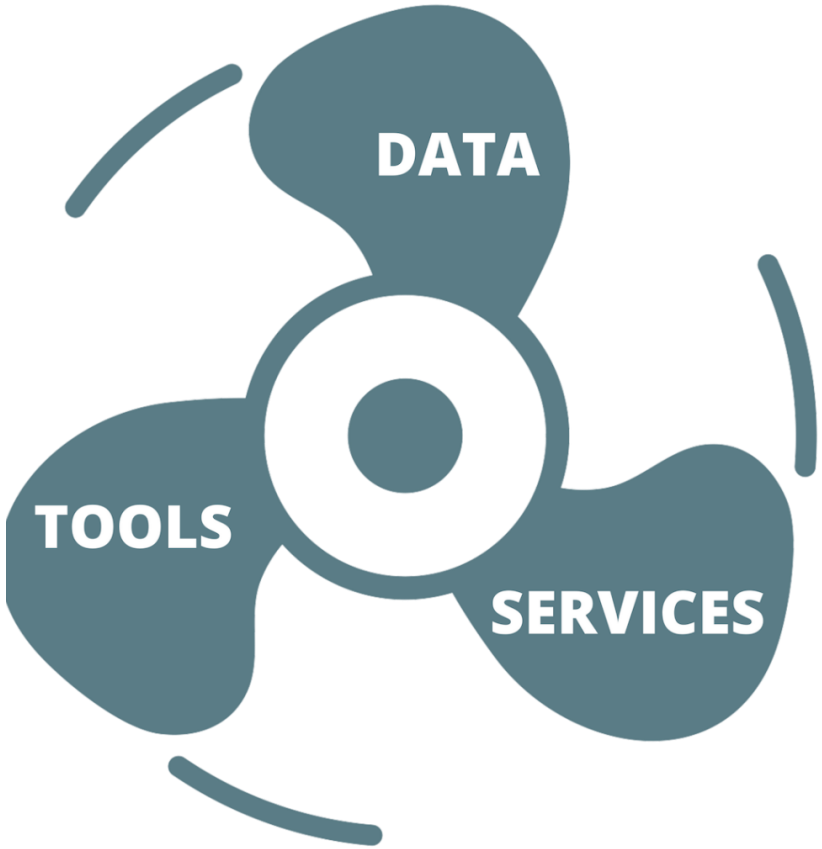
5ª Estrela - atribuída aos **dados que são conectados** (Linked Data - LD) a outros dados. Permite ampliar o contexto e a descoberta de informações.



BERNERS-LEE, T. Open, Linked Data for a Global Community. Gov 2.0 Expo. Washington. 2010. (http://5stardata.info/pt-BR/)

IFDS – INTERNET FAIR DE DADOS E SERVIÇOS

- É uma ideia de internet baseada no 'roteamento' escalável e transparente de dados, ferramentas e serviços (computacionais para executar as ferramentas)
- “Rotear dados de forma inteligente para ferramentas, ferramentas para dados e ambos para computação”
- Para além dos dados: “queremos tratar todos os 'objetos digitais' no IFDS de acordo com os mesmos princípios, incluindo a necessidade de metadados acionáveis por máquina suficientemente ricos” (GO FAIR, 2020)
- O conceito de ferramentas e serviços se intersectam: “Ferramentas - serviços de tipo de software que 'atuam sobre os dados', como, por exemplo, máquinas virtuais empacotadas para navegar pelo IFDS para análises de dados distribuídos.”

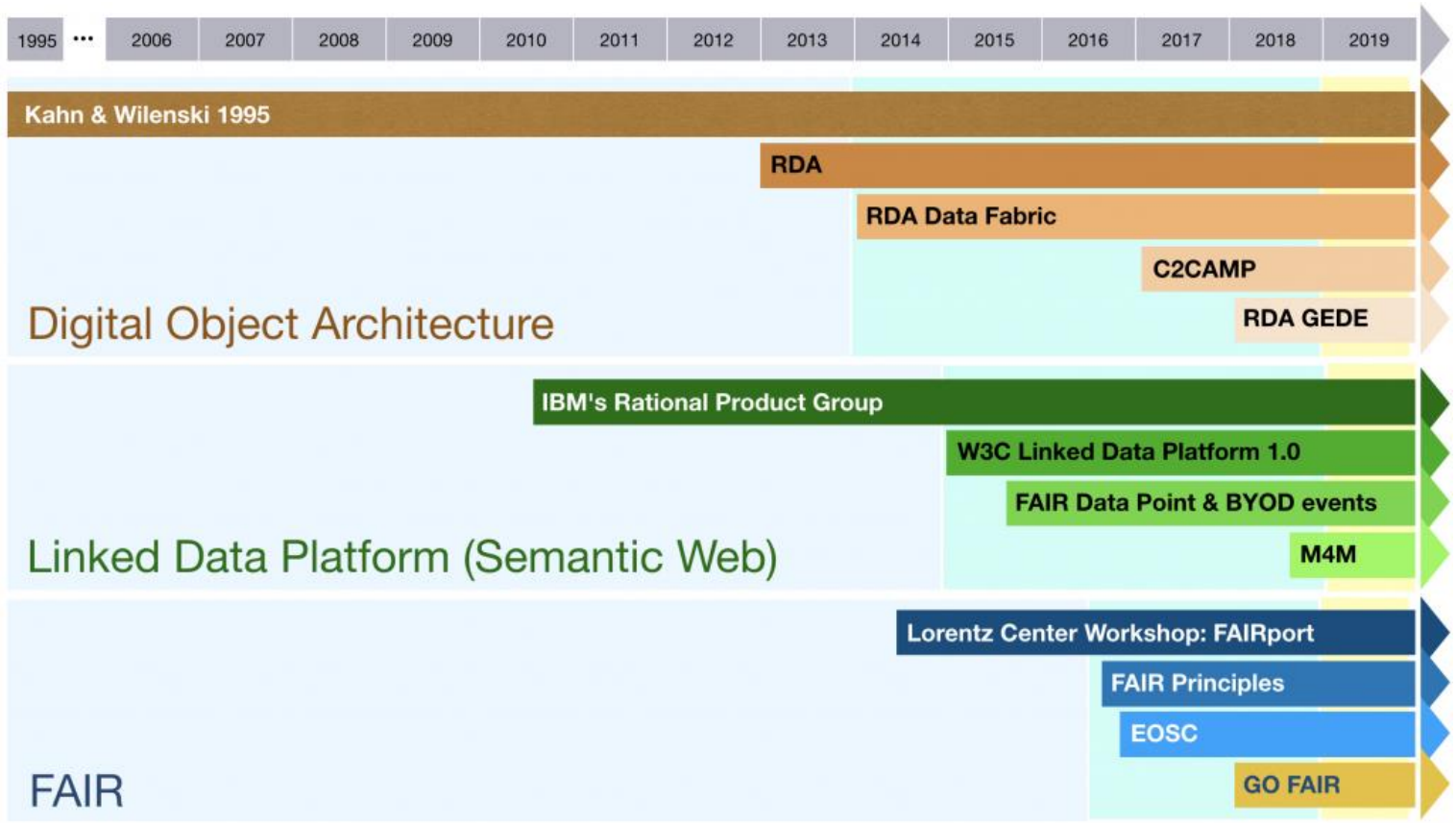


Dos Dados ao Objeto Digital

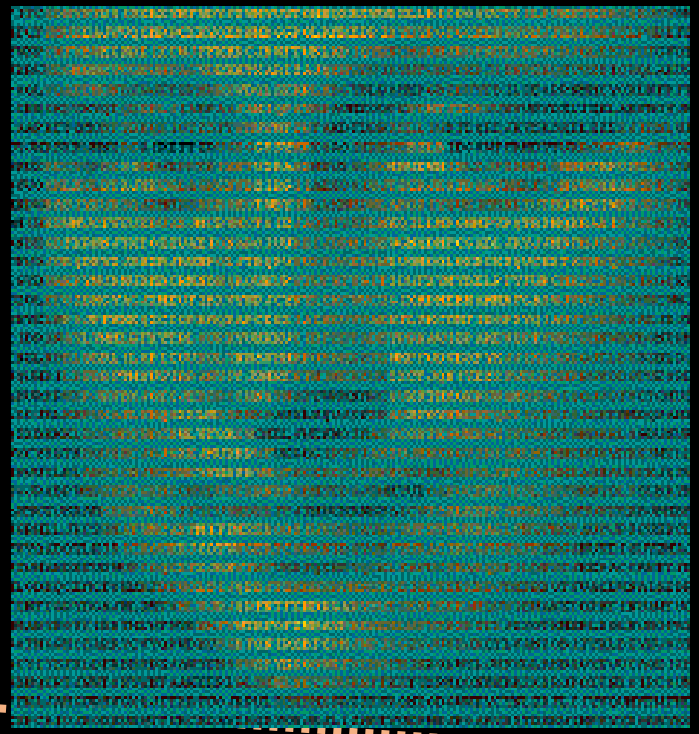
- Cada objeto digital é um contêiner composto por um referente, um identificador e um predicado (afirmação única legível por máquina) para um banco de dados ou software.

- Pré-requisitos para o IFDS:
  - A cada objeto digital seja atribuído um identificador único persistente e solvível (UPRI)
  - A adição específica do termo 'resolvível' aqui indica a necessidade de aceitar UPRIs múltiplos para apontar para o mesmo conceito.
  - Cada objeto digital deve ser descrito com **metadados rico o suficiente** em formato legível por máquina e seres humanos

Arquitetura de Objeto Digital



Timeline of developments in the convergence of various approaches to data infrastructures: conceptualization (blue background), design (green background), and implementation (yellow background)



OBJETOS DIGITAIS

Um objeto digital pode ser definido simplesmente como todo e qualquer objeto de informação que possa ser representado por meio de uma sequência de dígitos binários

TEXTO PRODUZIDO NO EDITOR DE TEXTO
ESSA APRESENTAÇÃO
FOTOGRAFIAS DIGITAIS
BASES DE DADOS
APLICACÓES DE SOFTWARE
MODELOS DE REALIDADE VIRTUAL
...

OBJETO DIGITAL NO CONTEXTO DOS SISTEMAS DE INFORMAÇÃO

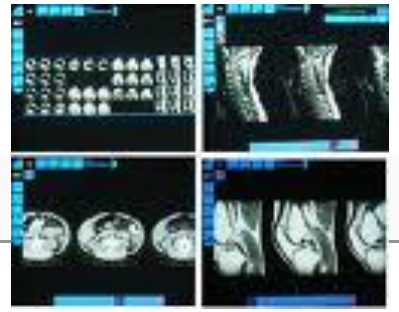
Itens armazenados em um repositório digital consistindo de dados, metadados e identificadores

Objetos digitais são conceitualmente equivalentes aos itens do acervo de bibliotecas, coleções de museus e documentos de arquivos
(NISO 2004)

Tipos de
OBJETOS DIGITAIS
Em relação à origem

Nascidos digitais

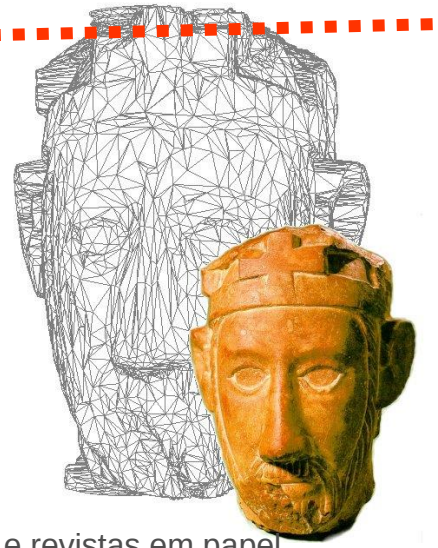
Já foram criados em formatos digitais;



Fotografias, áudio e vídeos digitais;
Essa apresentação;
Texto escrito em um editor de texto
Imagens de satélites
Imagens médicas

Digitalizados

Foram transformados de formas analógicas para formatos digitais; são **representações** de materiais que existem em formatos analógicos



Fotografias, livros e revistas em papel que sofreram processos de digitalização

Tipos de

OBJETOS DIGITAIS

Em relação a sua composição

Um objeto digital pode estar completo em um **único arquivo**, ou consistir de uma **multiplicidade de arquivos** vinculados por *links* (exemplo: página HTML), ou múltiplos arquivos.

Objetos digitais simples

Estão completos em único arquivo

Documento Word

Imagem JPEG

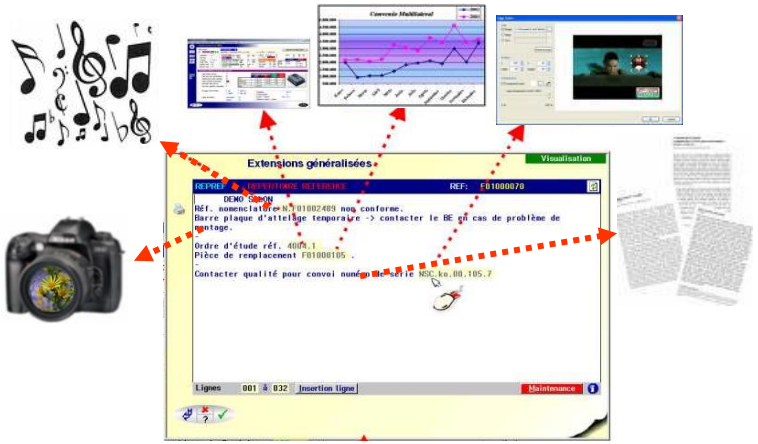
Objetos digitais complexos

Formados por um conjunto de arquivos e de metadados.

É visto como um único objeto conceitual

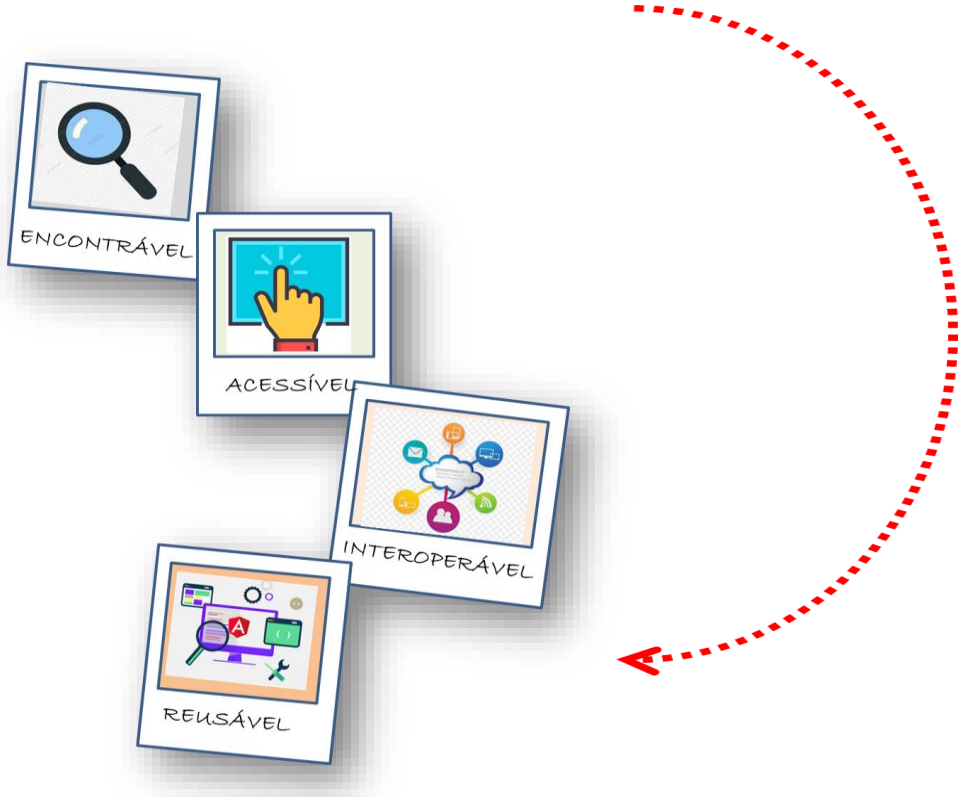
Um livro digitalizado por capítulos

Uma página Web

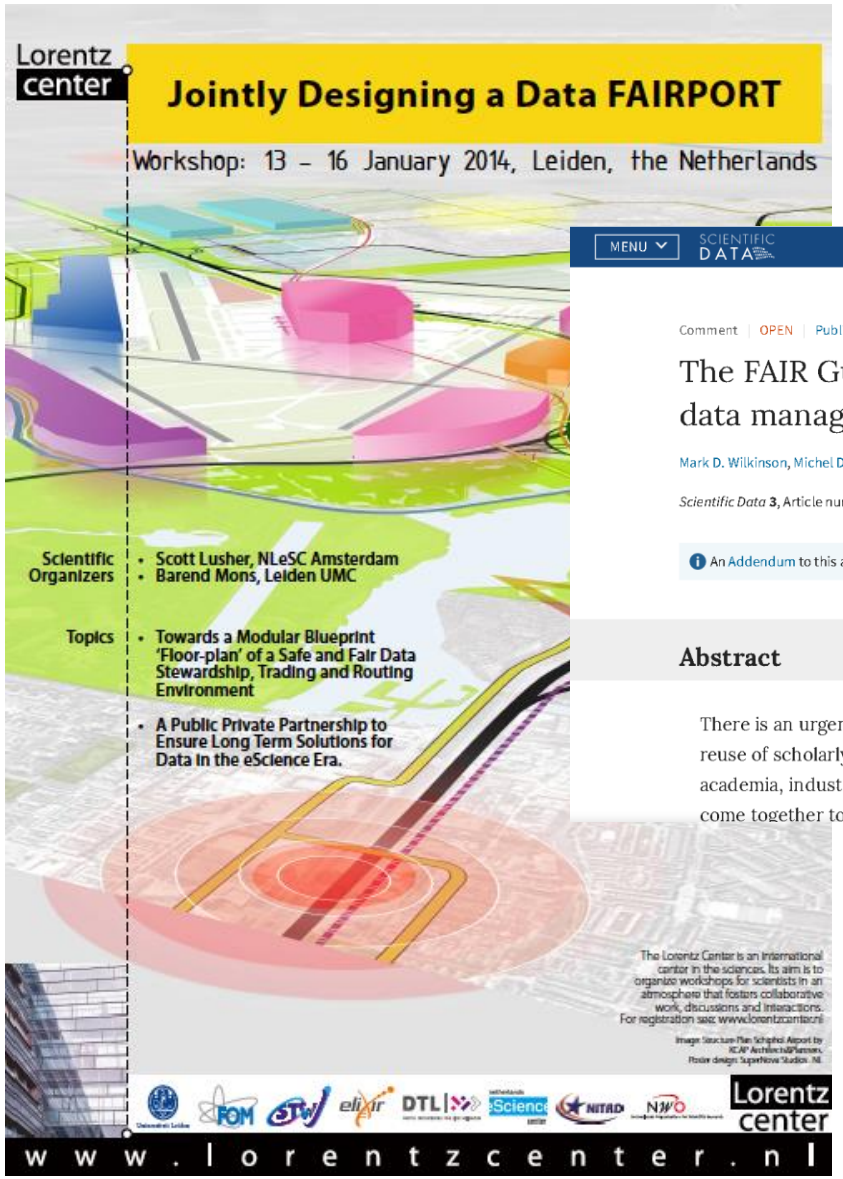


Afinal, o que são FAIR ?

- São aqueles que adotam quinze princípios criados para as melhores práticas de gestão, compartilhamento e reuso respeitando todas as questões éticas, legais e restrições contratuais.



- Criados no *Jointly designing a data FAIRPORT Conference,* por especialistas de diversas áreas do conhecimento interessados no reuso de dados, no contexto da e-Science em 2014
- Publicados em 2016



What is FAIR DATA?

Dados e materiais suplementares têm metadados suficientemente ricos e identificadores únicos e persistentes

FINDABLE

Metadados e dados são compreensíveis por homens e máquinas. Dados são depositados em repositórios confiáveis

ACCESSIBLE

Metadados usam uma linguagem formal, acessível, compartilhada e amplamente aplicada para representação de conhecimento.

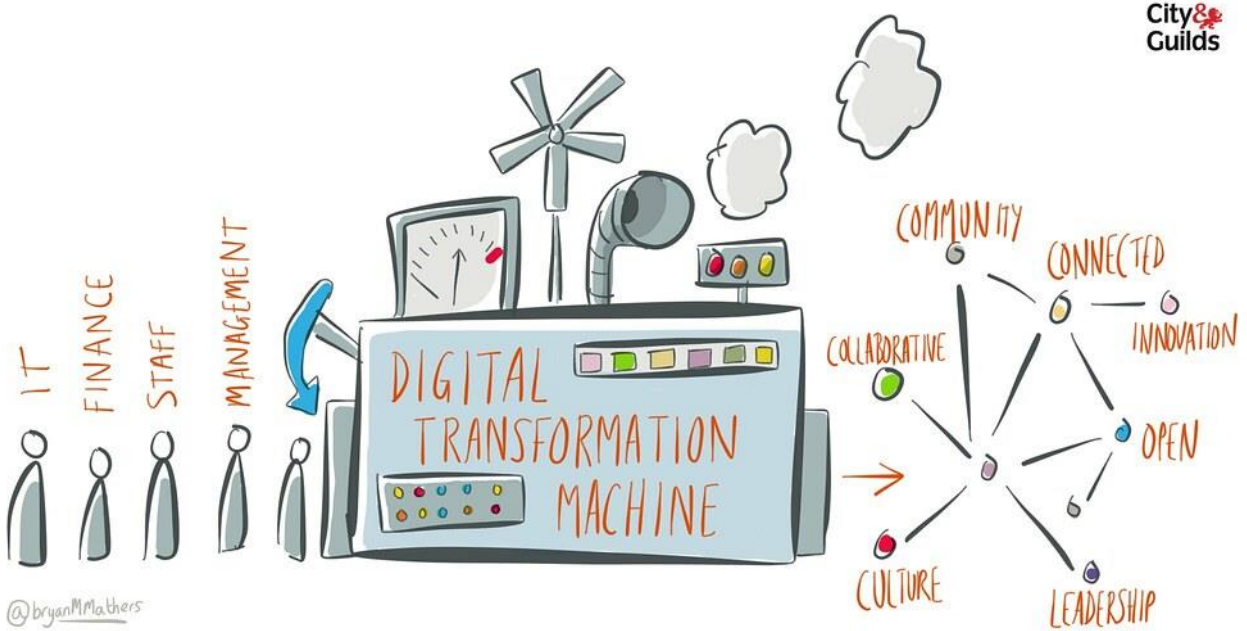
INTEROPERABLE

Dados e coleções têm uma licença de uso clara e dispõem de informação precisa sobre a proveniência

REUSABLE

Como tornar os dados FAIR ?

- Encontrável – dado/objeto deve ter identificador persistente único; e deve ser regisrado e indexado em um mecanismo de busca
- Acessivel – dado/objetos podem ser baixados por outros através de protocolos de comunicação padronizado
- Interoperável – objetos devem ser tratados com padrões abertos (formatos, lingaugens e vocabulários)
- Reusado - Devido a condições de uso conhecidas, o objeto poderá ser reusado por outros



Findable

Para apoiar a descoberta automática de datasets relevantes, (meta)dados precisam ser fácil de ser encontrados por humanos e por máquinas e ter um identificador persistente

Accesible

Limitação de uso de dados e protocolos para consultar ou copiar dados são explicitados tanto para humanos quanto para máquinas

Interoperable

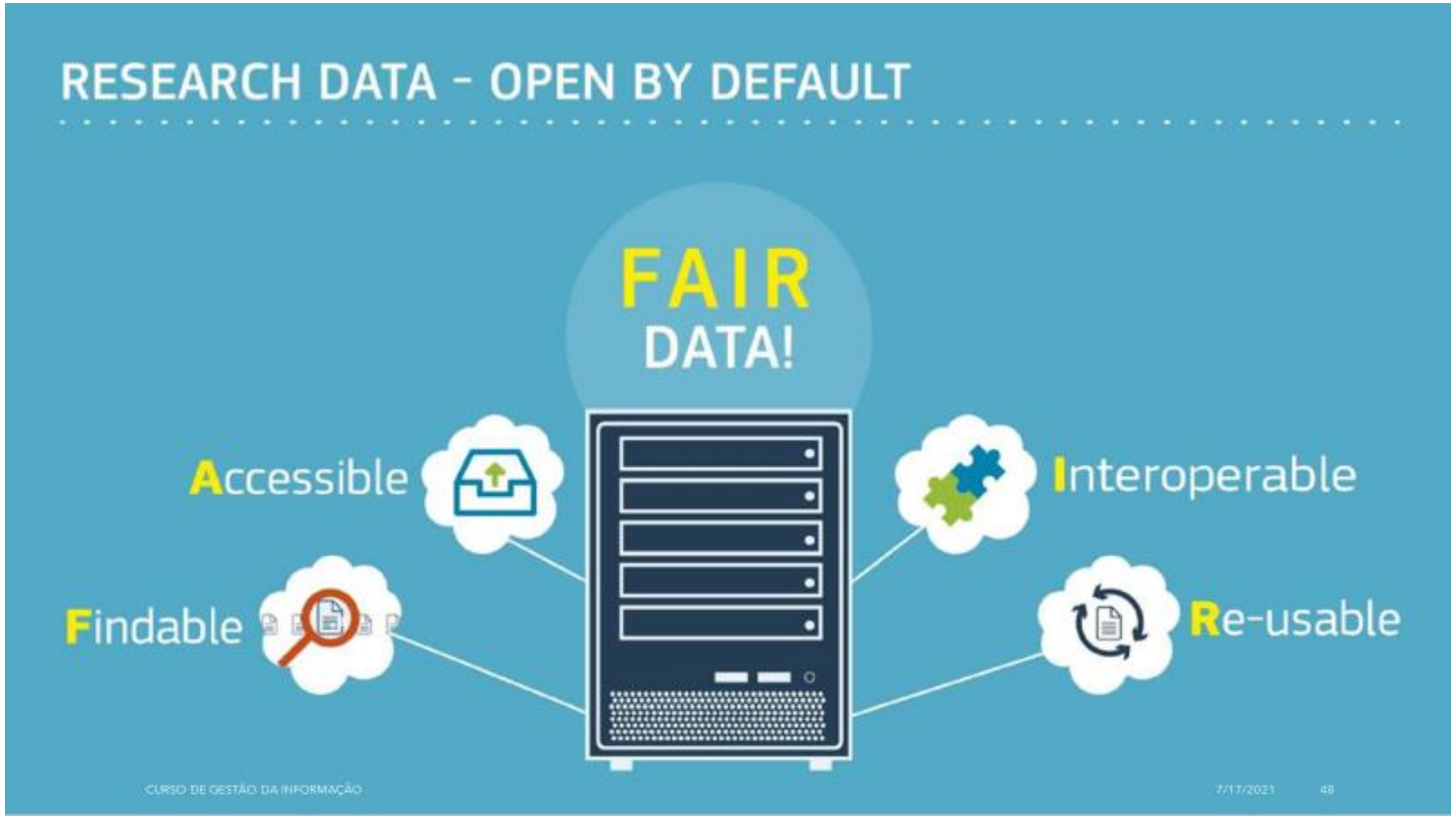
(Meta)dados devem usar termos padronizados (vocabulários controlados), referenciar outros (meta)dados e serem acionáveis por máquina

Reusable

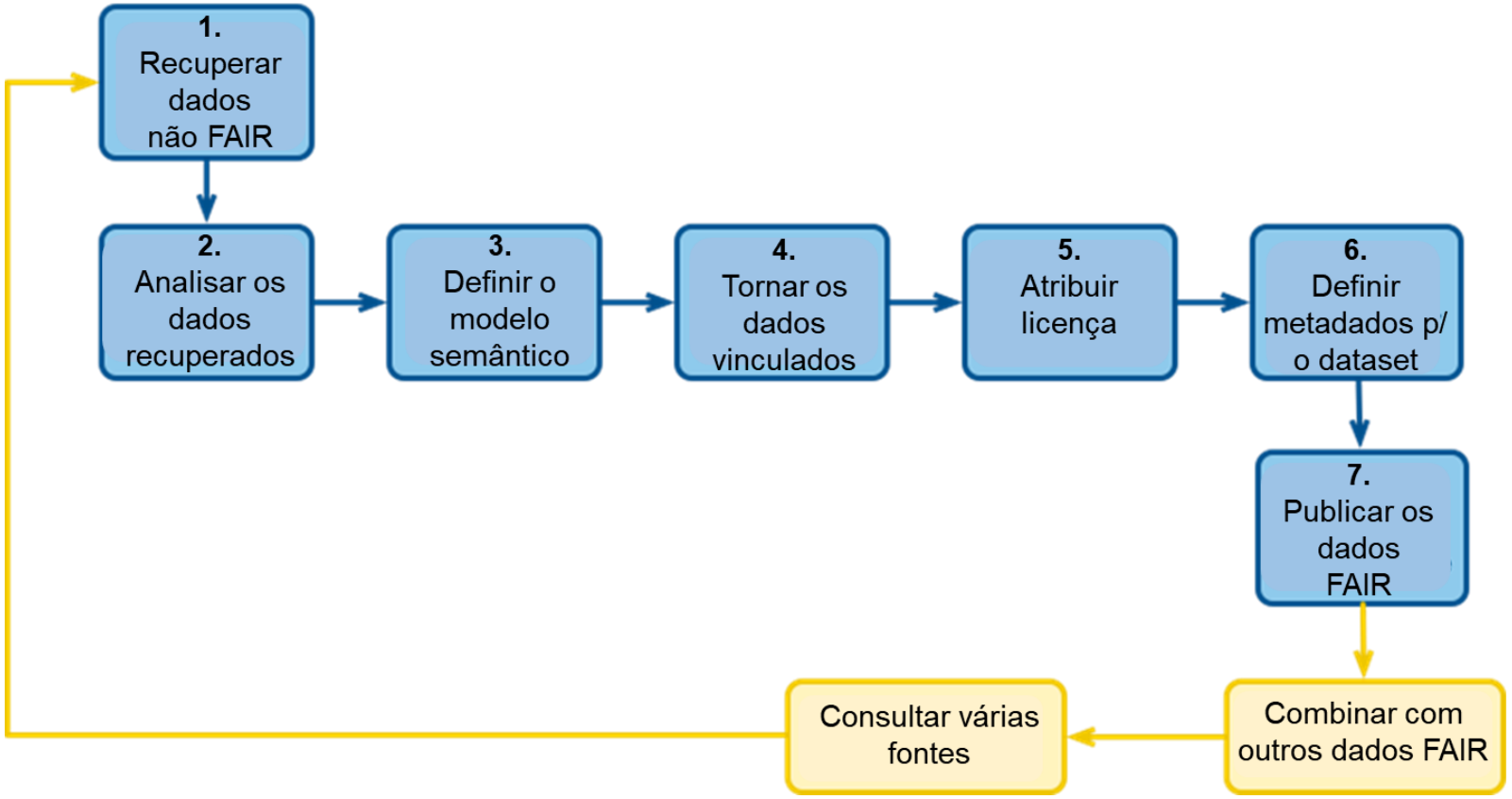
(Meta)dados são suficientemente bem descritos permitindo que eles sejam compreendidos por humanos e por máquinas e tenham uma licença de uso acessível e clara

Fairificação de dados

- Fairificar dados significa torná-los compatíveis com os princípios FAIR
- A maioria dos requisitos de localização e acessibilidade (F e A) aconteçam no âmbito dos metadados
- A interoperabilidade e o reuso (I e R), no nível dos dados.



Processo de FAIRificação (GO FAIR)

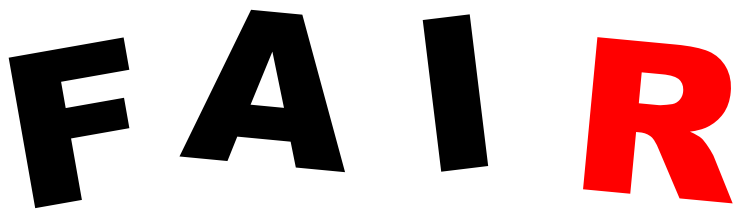


PASSOS PARA TORNAR SEUS DADOS FAIR

Encontrável

- Selecione um repositório e verifique os formatos de dados e metadados necessários
- Verifique se existe indendtificador persistente
- Selecione um catálogo pu mecanimso de busca que faça seu dado ser encontrado, especialmente se esse repositório for de natureza genérica

PRINCÍPIOS



Acessível

-Garanta longevidade aos dados (ex:tornando o repositório confiável através de certificação)

-Descreva as condições legais sob as quais os dados podem ser disponíveis e acessáveis

_Estabeleça embargo quando necessário

Esteja certo que sua infraestrutura materá o dado disponível no caso de defeito do equipamento ou erro humano.

Passos para tornar seus dados FAIR

Interoperável

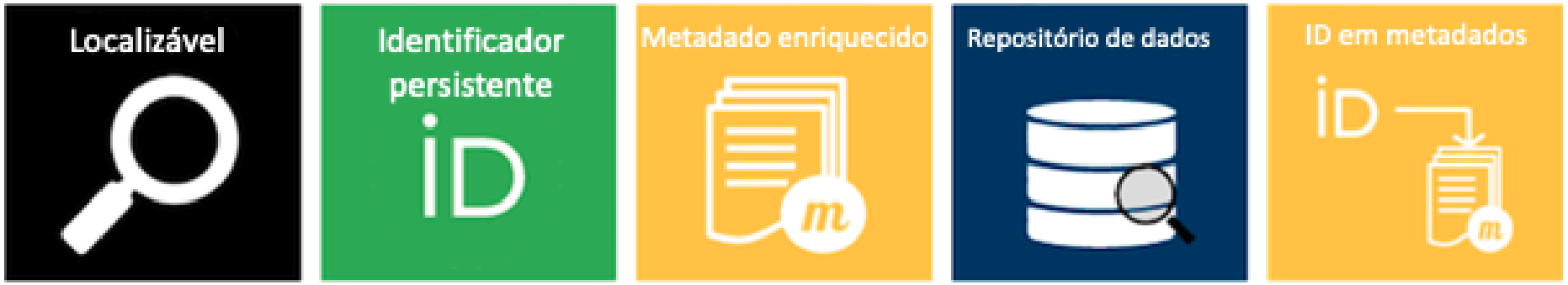
Selecione os formatos de dados mais usados
Selecione padrões mais usados
Selecione e aplique os vocabulários mais usados

Reusável

Registre informação sobre a proveniência do dado
Selecione padrões gerais mínimos de metadados
Atribua uma licença específica aos dados
Relacione outros documentos e informações sobre o assunto do dado

Encontráveis

- Os dados devem ser fáceis de serem encontrados.
- **Metadados** ricos devem estar disponiveis em uma ferramenta de busca online
- Devem ser associados a um identificador persistente (DOI)



F1. (meta)dados devem ter identificadores globais, únicos e persistentes

Adotar identificador único persistente tanto para o conjunto de dados quanto para os metadados (ex: DOI, handle, orcid, PID)

Identificação Persistente

Lembra a situação de um livro numa grande biblioteca que não está na estante na posição indicada no catálogo.

Como encontrá-lo?

Hoje o URL – Uniform Resource Locator - é a porta de entrada para os recursos que estão disponíveis na Web, ele define, como seu próprio nome diz, a localização do recurso.

erro 404

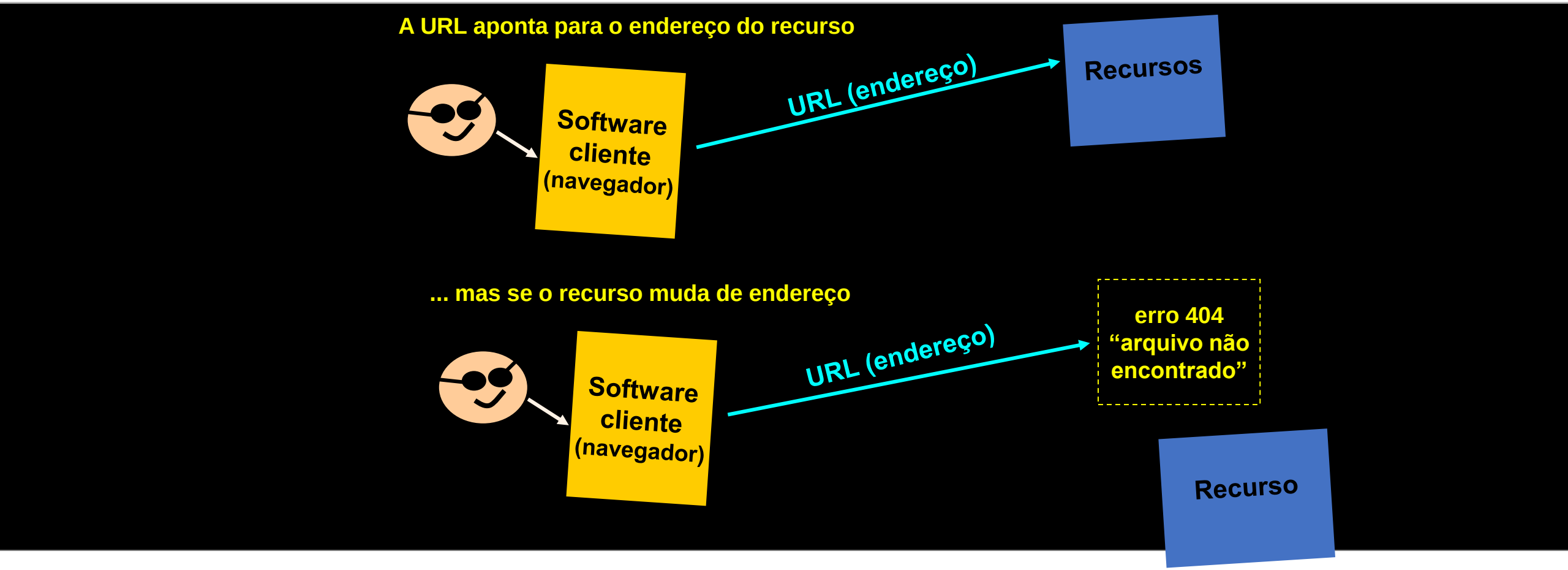
arquivo não encontrado

Freqüentemente tratamos o URL – Uniform Resource Locator – como se ele fosse um identificador. Na realidade, o URL é simplesmente um endereço mascarado como um identificador.

Confiar no URL como um identificador único para os recursos digitais, é como usar o endereço residencial no lugar do CPF

IDENTIFICADORES PERSISTENTES

Os objetos digitais precisam ser identificados de forma persistente



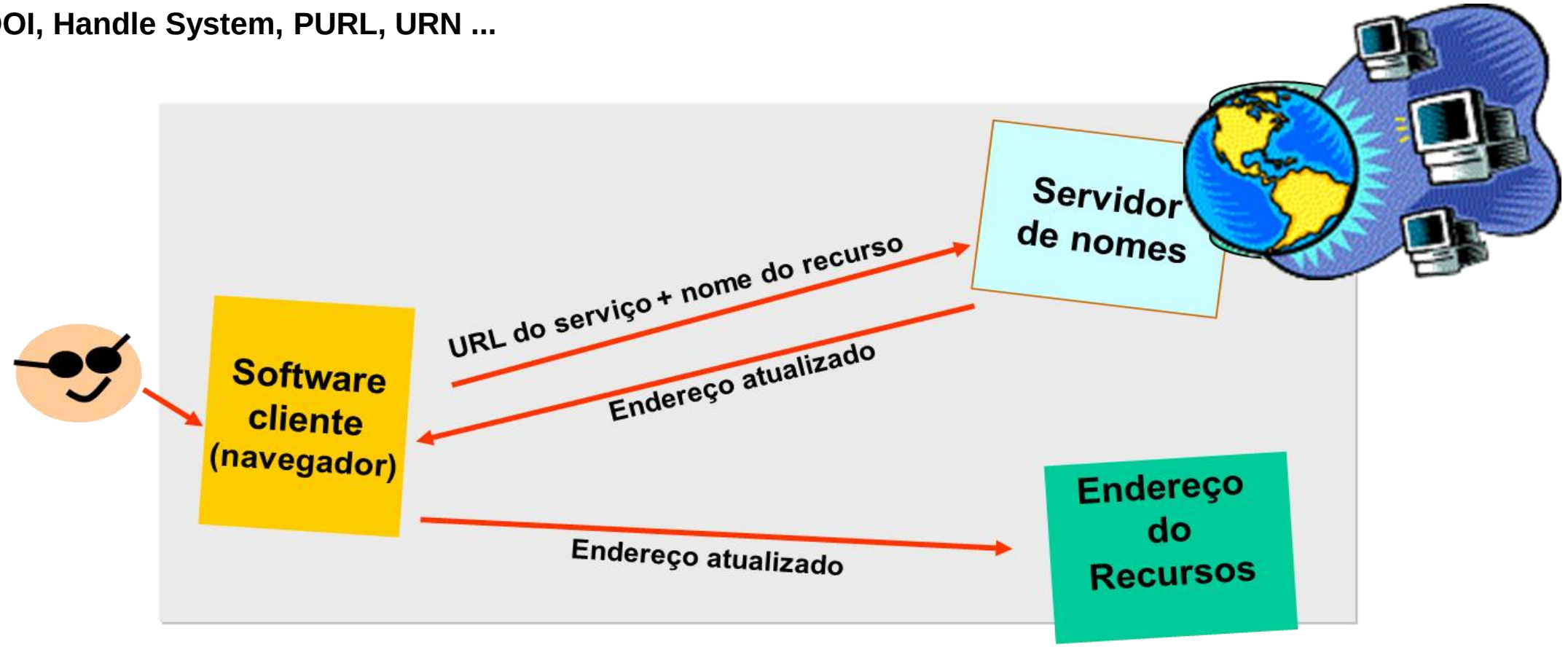
A URL é um endereço e não um identificador !

IDENTIFICADORES PERSISTENTES

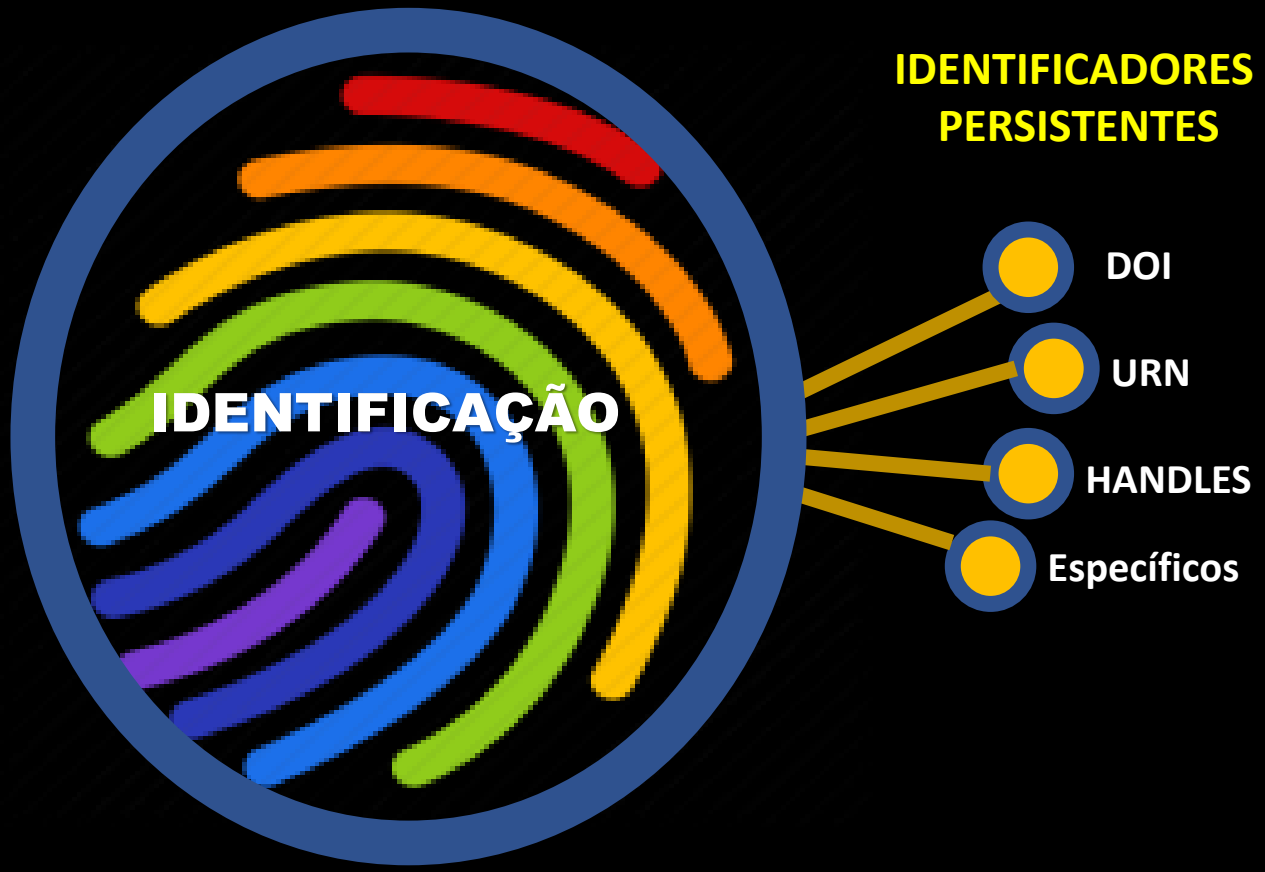
Nomes devem ser:
UNICOS, GLOBAIS, PERSISTENTES, INDEPENDETES DE LOCALIZAÇÃO E TECNOLOGIA, PADRONIZADOS

Para isto funcionar é necessário estabelecer uma **infraestrutura administrativa** para decidir quem pode assinalar nomes que identificam univocamente os recursos digitais de forma persistente.

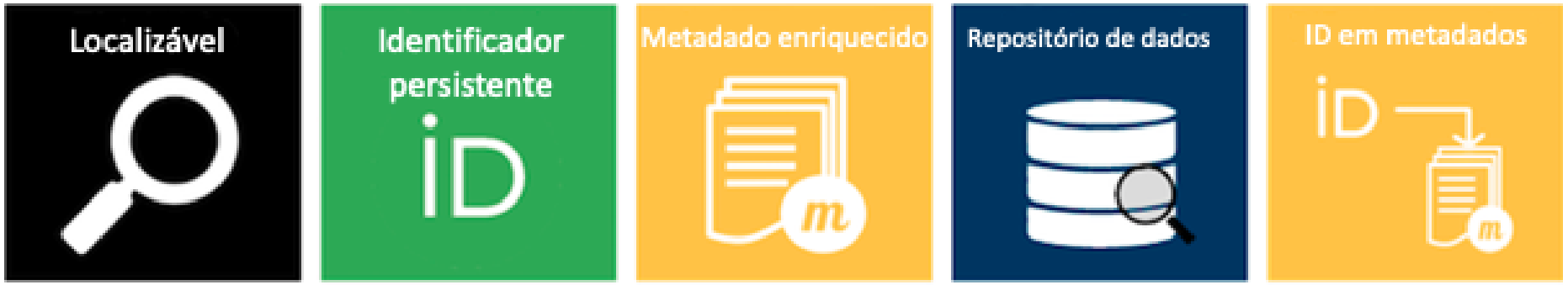
DOI, Handle System, PURL, URN ...



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*.

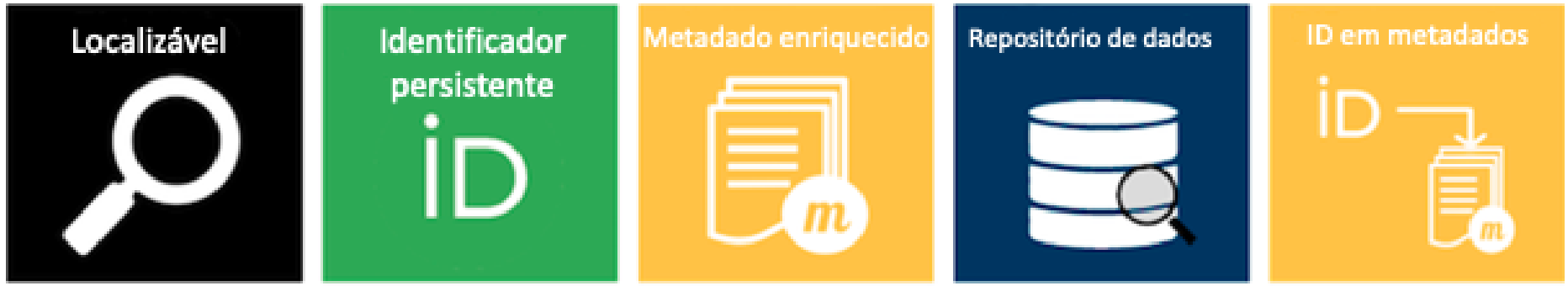


ACESSO
CITAÇÃO
PRESERVAÇÃO
LINK
INTEROPERABILIDADE



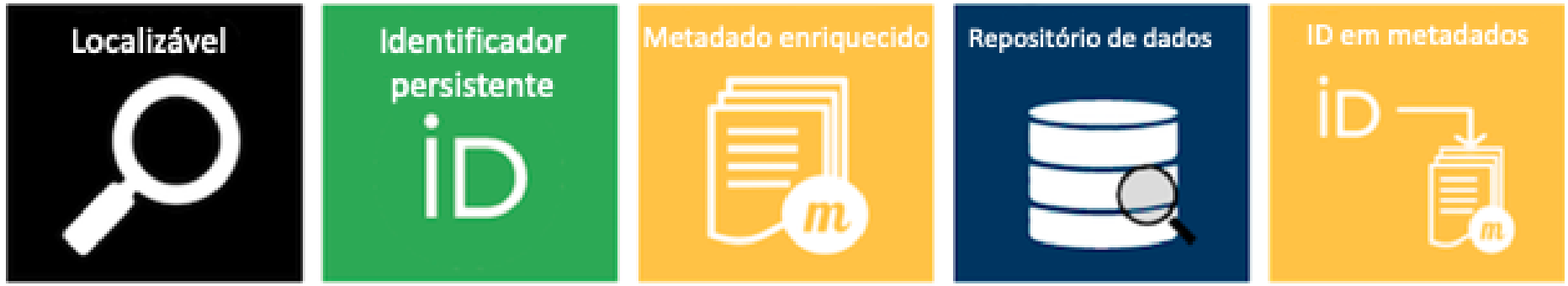
F2. dados devem ser descritos utilizando metadados ricos (impacta diretamente R1)

O conjunto de dados deve ser descrito por metadados ricos o suficiente para que, uma vez indexados em um mecanismo de busca, possam ser encontrados mesmo sem o seu identificador único persistente



F3. metadados devem incluir clara e explicitamente os identificadores dos dados que descrevem

Como não podemos prever que os dados e seus metadados estejam sempre juntos, a associação entre eles deve ocorre pela inclusão do identificador persistente dos dados nos metadados.



F4. (meta)dados devem ser registrados ou indexados em mecanismos de busca

Para que os dados sejam encontrados, seus metadados devem ser indexados em mecanismos de busca (search engine), que possibilitem aos computadores e usuários encontrá-los com facilidade.

ACESSIVEL

Homens e máquinas devem ter acesso aos dados sob condições específicas ou restritas, quando for apropriado.

Tão aberto quanto possível.
Tão fechado quanto necessários

≠

Open

A1. (meta) dados devem ser recuperáveis pelos seus identificadores usando protocolo de comunicação padronizado

Com o identificador persistente do conjunto de dados e/ou de seus metadados, o usuário poderá recuperá-los mais facilmente por meio de protocolos de comunicação padronizados. (ex: HTTP ou FTP)

A1.1 o protocolo deve ser aberto, gratuito e universalmente implementável

Independente de licenciamento dos dados e dos metadados, o protocolo de comunicação usado para dar acesso a eles deve ser aberto, gratuito e passível de ser implementado por qualquer interessado. (ex: HTTP ou FTP)

A1.2. o protocolo deve permitir procedimentos de autenticação e autorização, quando necessário

Dependendo das restrições de acesso aos dados e/ou metadados, um mecanismo de autenticação e autorização para o acesso deve ser liberado pelo protocolo de comunicação. (Ex: os repositórios confiáveis oferecem essa opção)

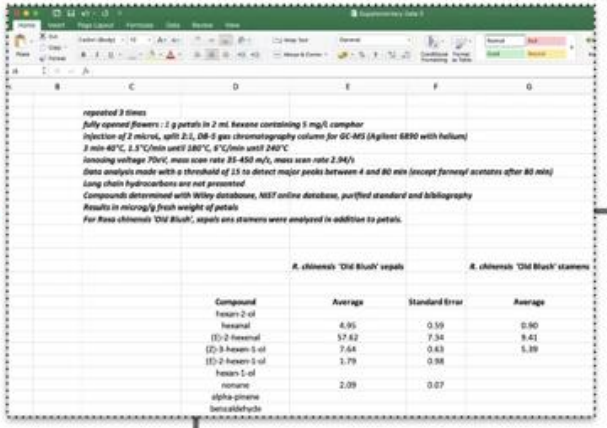
A2. metadados devem ser acessíveis, mesmo quando os dados não estiverem mais disponíveis.

É preciso existir um conjunto de estratégias de preservação para dados e metadados. Os metadados devem ser sempre acessíveis, possibilitando a criação de índices para o conjunto de dados atuais vigentes e aqueles não mais disponíveis.

- “As duas primeiras categorias (Findable e Acessible) se referem a processos que tornam os dados significativos para que as duas últimas categorias (Interperable e Reusable) se tornem possíveis. Isto é, dados só são interoperáveis e reusados (por máquinas ou seres humanos) se forem encontrados e acessados e para isso é necessário que o tratamento/curadoria desses dados seja realizado com base em padrões bem estabelecidos, seja no nível da descrição sintática, seja no nível da descrição semântica.”(SALES, 2020)

FAIR Data for Humans and Machines

START
Human-understandable, ambiguous data



Findability, Accessibility
Publish and get persistent resolvable identifier

Step 1

DOI 10.5281/zenodo.2598799

Interoperability
Transformation to an open syntax

Step 3

Step 2

Interoperability
Semantic mapping

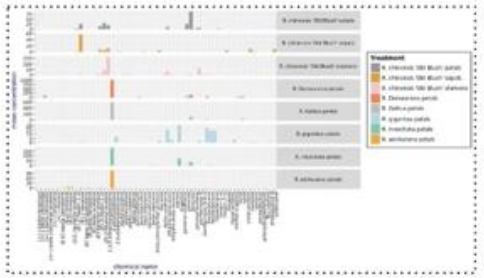
stato statistics ontology

ChEBI NCBI Taxonomy Plant Ontology

FRICTIONLESS DATA
SPECIFICATIONS AND SOFTWARE

Step 4

Reusability
Creation and querying of RDF Linked Data



Step 5

DOI 10.5281/zenodo.2640873

Findability, Accessibility
Publish and get persistent resolvable identifier

Step 5

DOI 10.5281/zenodo.3560778

RESULT
Human & Machine understandable, unambiguous data

INTEROPERÁVEIS

Os dados e os metadados devem estar em conformidade com formatos e padrões reconhecidos para permitir que sejam combinados e trocados entre sistemas.

INTEROPERÁVEL

I1. (meta) dados devem ser representados por meio de uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento

Para que se possa representar dados e metadados devem ser adotadas linguagens de representação do conhecimento que sejam padronizadas, acessíveis e amplamente aplicáveis. (Ex: RDF, XML, DICOM, etc.)

INTEROPERÁVEL

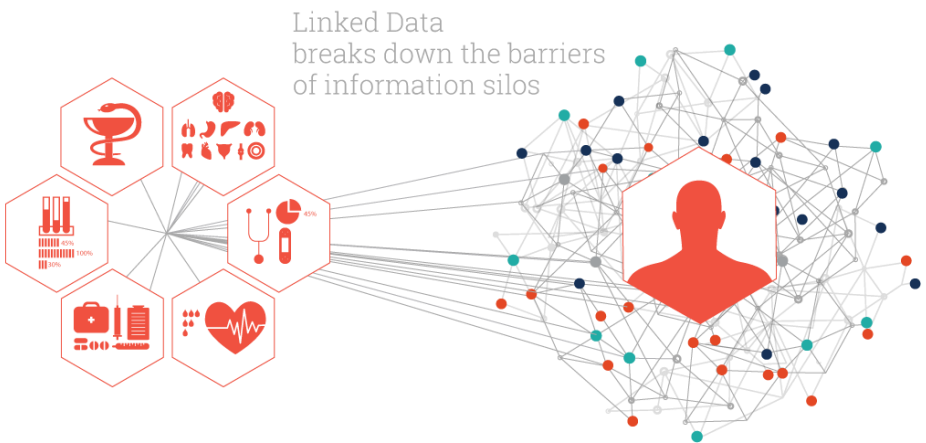
I2. (meta) dados devem usar vocabulários de acordo com os princípios FAIR

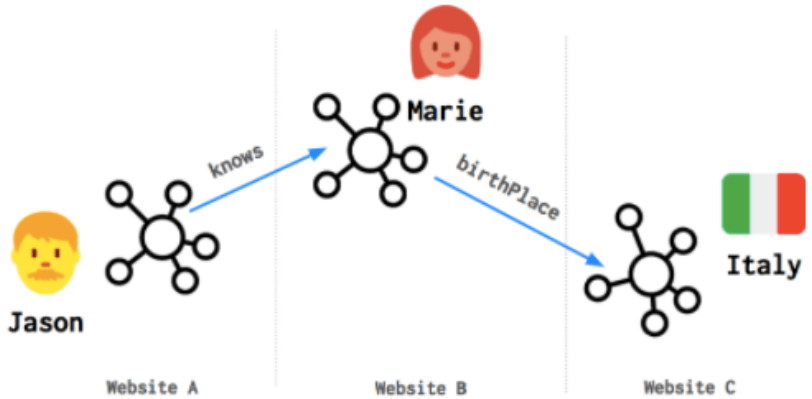
Dados e metadados devem possuir referências a vocabulários e/ou ontologias que os descrevem. Devemos garantir que esses também sigam os princípios FAIR.

INTEROPERÁVEL

I3. (meta) dados devem incluir referências qualificadas para outros (Meta) dados

É necessário referenciar o conjunto de dados, possibilitando que aqueles gerados a partir de outros conjuntos, sejam interligados. Assegurando a ligação semântica entre eles.





REUSÁVEIS

É necessário documentação apropriada para apoiar a interpretação e reutilização de dados.

Os dados devem estar em conformidade com as normas da comunidade e ter uma licença clara para que outros saibam quais tipos de reuso são permitidos.

R1. (meta) dados são descritos com uma pluralidade de atributos precisos e relevantes.

Prover metadados com alto nível de detalhes que permita ao pesquisador avaliar a possibilidade do seu reuso bem como adequação às suas necessidades.

R1.1. (meta) dados devem ser disponibilizados com licenças de uso claras e acessíveis

É fundamental que o responsável pelos dados e metadados defina explicitamente quem pode ter acesso a eles, com que finalidade e sob quais condições. Essas informações são definidas por meio de **licenças de uso.**

https://opendatacommons.org/

- Licença Open Data Commons Open Database (ODbL)
- Licença de atribuição Open Data Commons
- Dedicação e Licença de Domínio Público Open Data Commons (PDDL)

LICENÇAS PARA GESTÃO DE DADOS

O grupo Open Data Commons (http://opendatacommons.org/) tem desenvolvido ferramentas juridicamente vinculativas para controlar o uso de conjuntos de dados. Usando uma combinação de direitos autorais e padrões contratuais, eles criaram três licenças padrão que podem ser usadas em conjunto com projetos de dados. Além disso, é possível articular um conjunto de "normas comunitárias" que complementam o uso de licenças formais. Embora não tenham força de lei, as normas podem expressar as crenças compartilhadas de uma comunidade em relação ao compartilhamento e reutilização de dados.

As três licenças ODC são:

1. Dedicação e Licença de Domínio Público (PDDL): Dedica o banco de dados e seu conteúdo ao domínio público, livre para que todos possam usar como quiserem.
2. Licença de atribuição (ODC-By): os usuários são livres para usar o banco de dados e seu conteúdo de maneiras novas e diferentes, desde que forneçam atribuição à fonte dos dados e / ou ao banco de dados.
3. Licença de banco de dados aberto (ODC-ODbL): ODbL estipula que qualquer uso subsequente do banco de dados deve fornecer atribuição, uma versão irrestrita do novo produto deve estar sempre acessível e quaisquer novos produtos feitos usando material ODbL devem ser distribuídos usando os mesmos termos. É a mais restritiva de todas as licenças ODC.

LICENÇAS PARA GESTÃO DE DADOS

O Creative Commons (http://www.creativecommons.org) também possui uma biblioteca de licenças padronizadas, e algumas delas podem ser aplicadas a dados e bancos de dados. A licença ODC-By, por exemplo, é equivalente a uma licença Creative Commons Attribution (CC BY). As licenças CC BY, no entanto, são baseadas na propriedade dos direitos autorais da obra subjacente, enquanto a licença ODC-By pode se aplicar a obras que não são protegidas por direitos autorais (como dados factuais). As três licenças CC de maior relevância para o gerenciamento de dados são:

1. CC0 (ou seja, "CC Zero"): Quando um proprietário deseja renunciar aos seus direitos autorais e / ou direitos de banco de dados, a marca CC0 pode ser usada. Ele efetivamente coloca o banco de dados e os dados em domínio público. É o equivalente funcional de uma licença ODC PDDL.

2. Marca de domínio público (PDM): é usada para marcar trabalhos que são de domínio público e para os quais não há direitos autorais conhecidos ou restrições de banco de dados. Dados factuais em um banco de dados, por exemplo, podem ser marcados como PDM para deixar claro que seu uso é gratuito

1. CC-BY : É usado quando um proprietário deseja permitir que seu trabalho protegido por direitos autorais seja reutilizado e compartilhado com a condição de que o crédito apropriado seja dado.

R1.2. (meta) dados devem estar associados à sua proveniência

- Especificar a **proveniência** (linhagem) dos dados é importante não só para que o pesquisador possa avaliar a utilidade dos dados ou metadados, mas também para que possa atribuir o devido crédito a quem produziu, manteve ou editou esses dados.
- Dentre as informações relativas à proveniência destacam-se:
  - A **linhagem** dos dados, ou seja, o processo de obtenção dos dados (gerado ou coletado);
  - Particularidades ou limitações sobre os dados que outros pesquisadores devem conhecer;
  - **Data da geração** do conjunto de dados, **condições** de laboratório, **quem** preparou os dados, **configurações** de parâmetros, nome e versão do **software** utilizado;
  - Explicitar o nível de processamento dos dados (são dados brutos ou processados?);
  - A **versão** dos dados arquivados e/ou reutilizados deve ser claramente especificada e documentada.

R1.3. (meta) dados devem estar alinhados com padrões relevantes do seu domínio

Além de atender aos padrões específicos de cada comunidade deve-se dar atenção às boas práticas de arquivamento e compartilhamento específicos da área de pesquisa.

Em suma...

ACIONALIDADE POR MÁQUINA

- Os seres humanos não são os únicos interlocutores críticos no ecossistema de dados, o **FAIR é principalmente para máquinas;**
- Escopo, escala e velocidade requisitada pelo nível de complexidade da ciência contemporânea;
- Os computadores devem ser capazes de acessar os dados de forma autônoma;
- Devem encontrar e usar dados e apoiar o reuso por humanos;
- Os **"stakeholders computacionais"** são exploradores que agem em nosso nome: agentes, programas de aplicação etc.

FAIR é sobre...

METADADOS

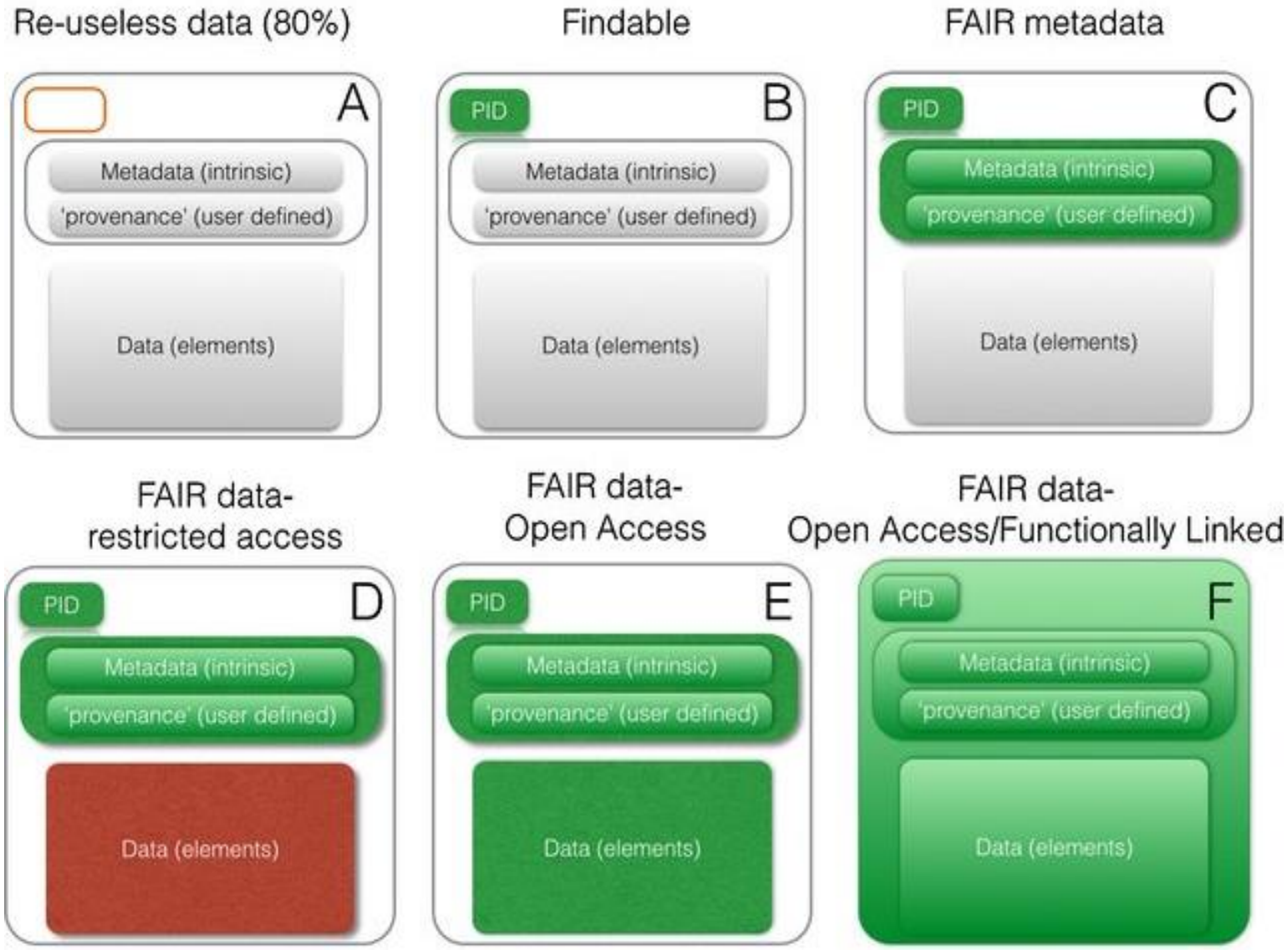
- A acionalidade por máquina coloca em destaque **a importância dos metadados** que estão presentes nos 15 princípios.
- O objeto digital deve fornecer informações cada vez mais detalhada para um explorador computacional;
- O que é o objeto, contexto, estrutura e intenção, utilidade no contexto, licença, consentimento, nível de sensibilidade;

ACESSO SOB CONDIÇÕES BEM DEFINIDAS

- **FAIR não é igual a aberto**;
- Há razões legítimas para blindar os dados e serviços gerados com fundos públicos que requerem medidas adicionais de autorização e autenticação tanto para humanos quanto para máquinas;
- FAIR não endereça questões e morais sobre a abertura dos dados: critério do custodiante.

POR QUE APLICAR PRINCÍPIOS FAIR AOS REPOSITÓRIOS

NIVEIS DE *FAIRNESS* X ABERTURA DOS DADOS



A
80% dos datasets está indisponível para o reuso;

B
Primeiro passo: PID – identificador persistente;

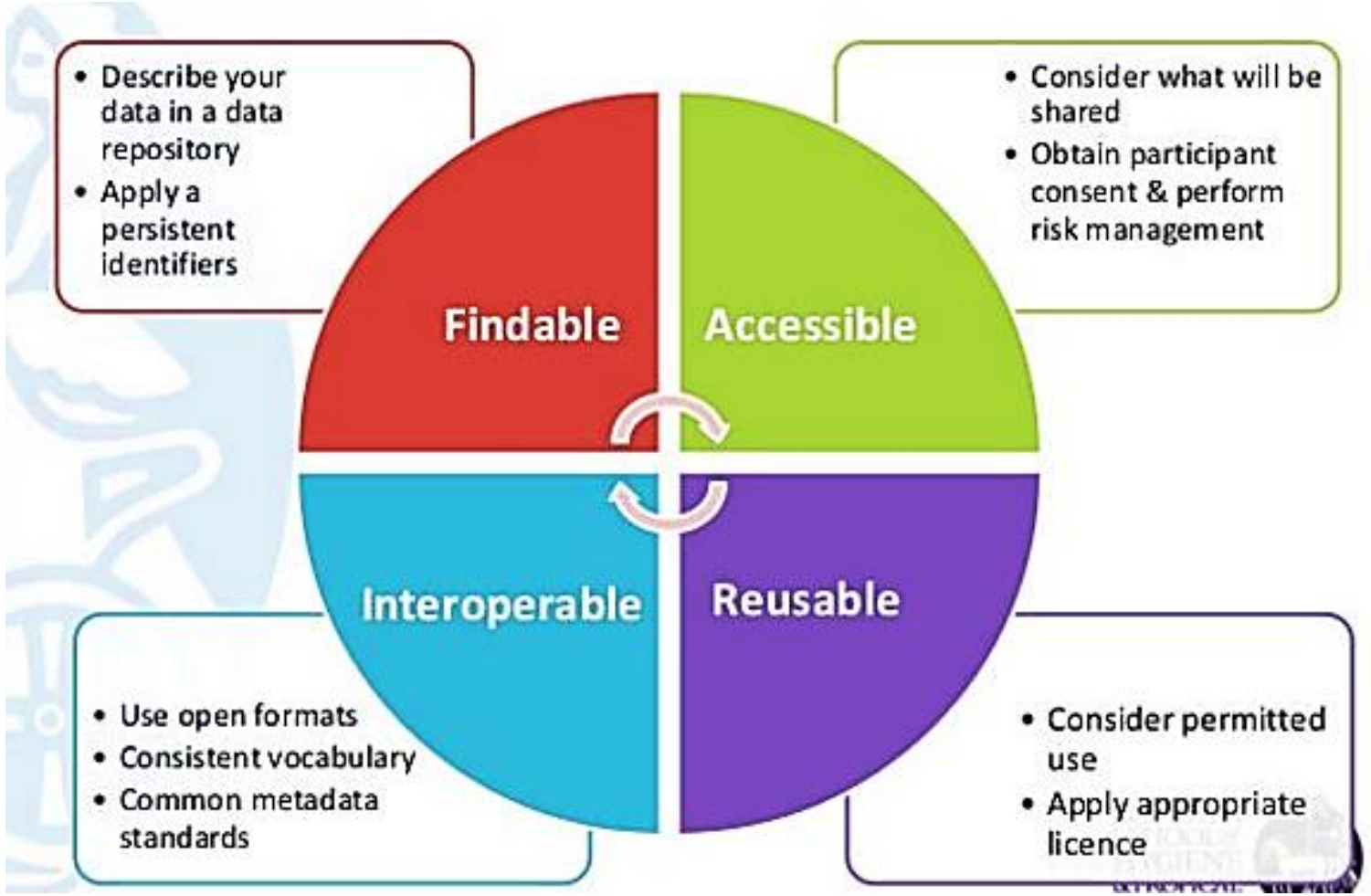
C
Metadados legíveis por máquina: intrínsecos e definidos pelo usuários (contexto e proveniência)

D
Dados com restrições de acesso

E
Dados disponíveis sob condições bem definidas para reuso.

F
Internet FAIR data: número de aplicações e serviços podem linkar e processar dados FAIR

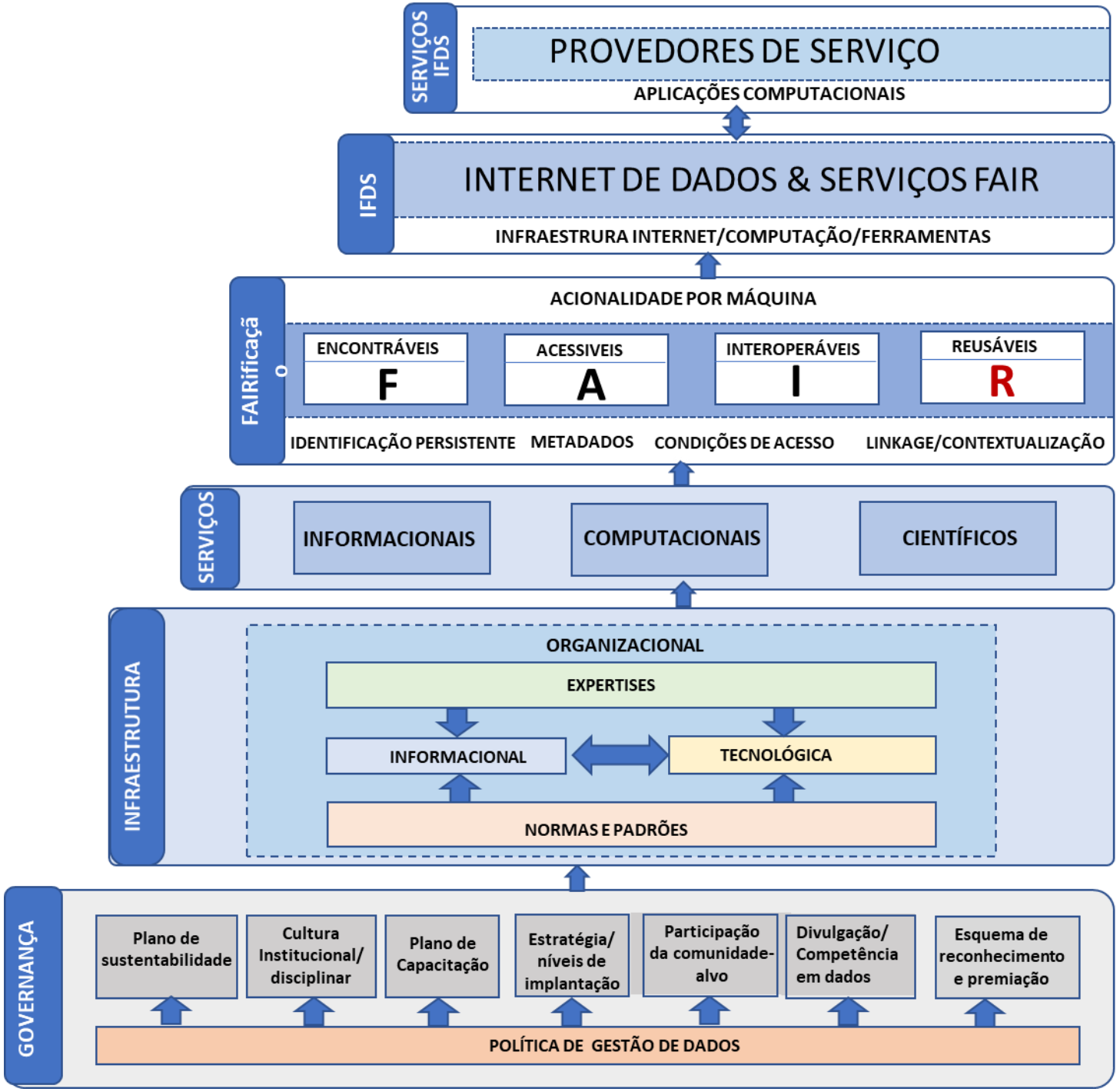
COMO TER UM REPOSITÓRIO FAIR ?

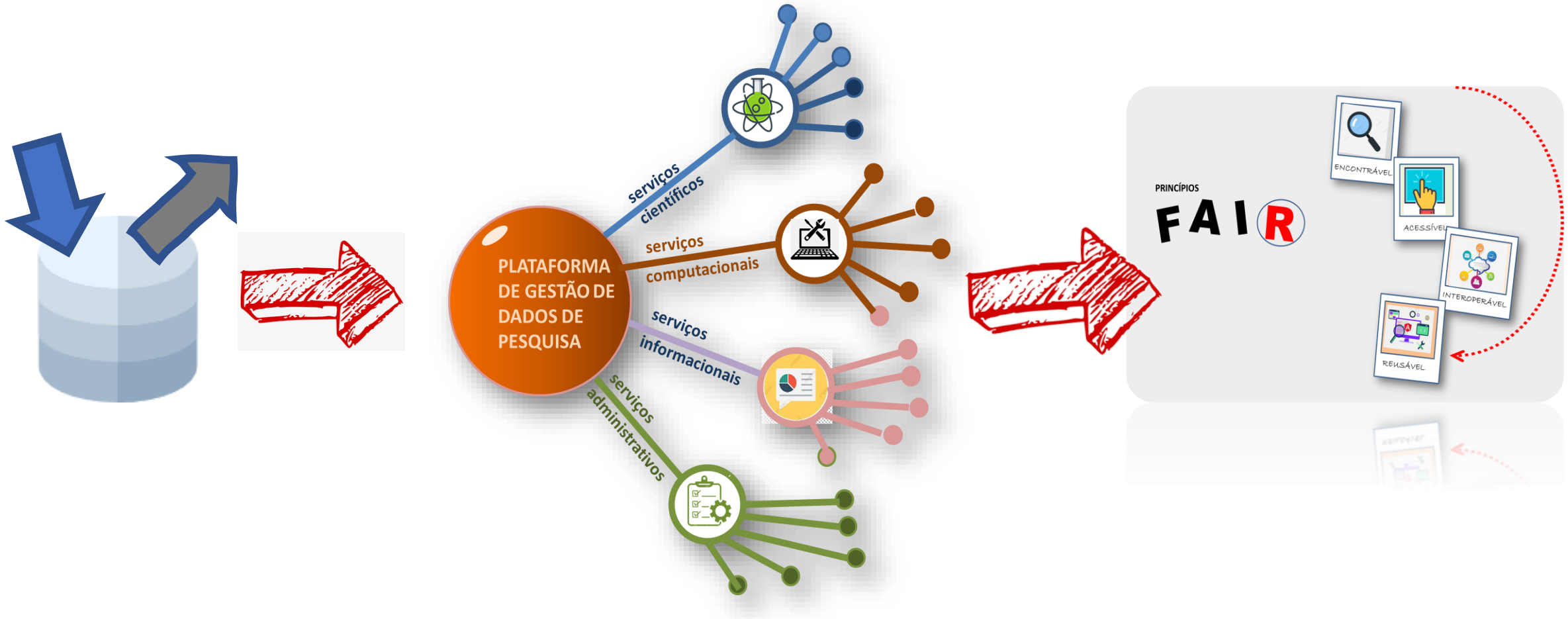


DADOS FAIR

Nos levará à

IFDS





REPOSITÓRIO

memória/depósito/acesso

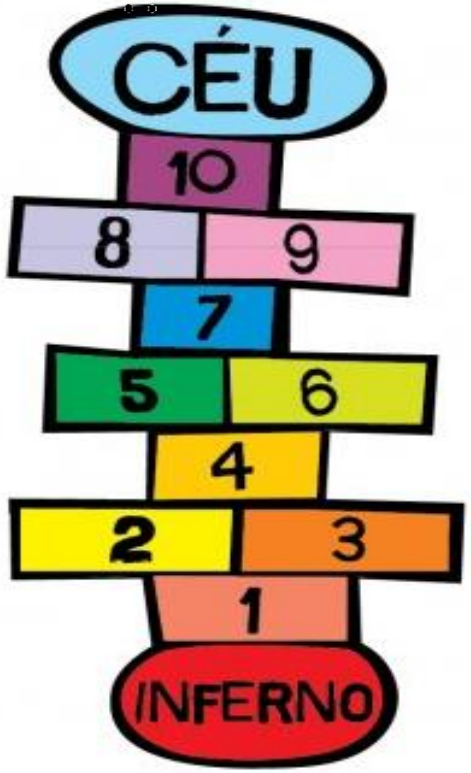
GESTÃO DE DADOS COMO SERVIÇO

Enfoque disciplinar/comunitario

FAIRIFICAÇÃO


SERVIÇOS DE GESTÃO DE DADOS DE PESQUISA
serviços científicos
serviços computacionais
serviços informacionais
serviços administrat

UMA ARQUITETURA PARA A GESTÃO DE DADOS COMO SERVIÇO



FAIRificação

ACIONALIDADE POR MÁQUINA

ENCONTRÁVEIS	ACESSIVEIS	INTEROPERÁVEIS	REUSÁVEIS
F	A	I	R

SRVIÇOS

CIENTÍFICOS

COMPUTACIONAIS

ANTES DA PESQUISA

DURANTE A PESQUISA

DEPOIS DA PESQUISA

INFORMACIONAIS

ADMINISTRTIVOS

INFRAESTRUTURA

ORGANIZACIONAL

EXPERTISES

INFORMACIONAL

TECNOLÓGICA

NORMAS E PADRÕES

GOVERNANÇA

Plano de sustentabilidade

Cultura Institucional/ disciplinar

Plano de Capacitação

Estratégia/ níveis de implantação

Participação da comunidade-alvo

Divulgação/ Competência em dados

Esquema de reconhecimento e premiação

POLÍTICA DE GESTÃO DE DADOS

GOVERNANÇA DE DADOS DE DADOS DE PESQUISA

Delineia os princípios, políticas e estratégias que são comumente adotados num ambiente que necessita de um programa de gestão de dados coerente; delineia também as ações, funções e papéis que são necessários para implementar essas políticas e estratégias

Este arcabouço estruturante é necessário posto que dados de pesquisa digitais só podem ser gerenciados e preservados adequadamente ao longo do tempo por meio de um compromisso institucional sustentado (MAYERMIK, 2012, p.1).

GOVERNANÇA DE DADOS DE DADOS DE PESQUISA

PLANO DE SUTENTABILIDADE

CULTURA INSTITUCIONAL E DISCIPLINAR

PLANO DE CAPACITAÇÃO

ESTRATÉGIA DE IMPLANTAÇÃO

PARTICIPAÇÃO DA COMUNIDADE

DIVULGAÇÃO/COMPETÊNCIA EM DADOS

ESQUEMAS DE PREMIAÇÃO

POLÍTICA DE GESTÃO DE DADOS

INFRAESTRUTURA

INFRAESTRUTURA ORGANIZACIONAL

TECNOLÓGICA

INFORMACIONAL

NORMAS E PADRÕES

EXPERTISES

Da mesma forma que as **instituições devem providenciar infraestruturas básicas para a pesquisa** – tais como laboratórios, instrumentação, computação de alto desempenho, redes, reagentes e muito mais – elas **devem também tomar medidas para uma gestão adequada dos dados**. Isto pressupõe um amplo espectro de atividades gerenciais, tecnológicas e informacionais que inclui profissionais de informação treinados para apoiar pesquisadores no planejamento e gestão de seus dados, no acesso a dispositivos de armazenamento seguro e *backups* durante o desenvolvimento do projeto e disponibilidade de plataformas de acesso e de preservação de longo prazo, necessárias após o fim da pesquisa (STRASSER, 2015); Os arcabouços infraestruturais voltados para a gestão de dados **são diversos e fragmentados** em termos de fluxos, complexidade, aplicação e topologia, e organizados de forma diferente pelas várias disciplinas e em diferentes países (GRAAF; WAAIJERS, 2011).


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

INFRAESTRUTURA

INFRAESTRUTURA ORGANIZACIONAL - O arcabouço infraestrutural pressupõe, assim como a governança, uma ancoragem baseada em alguma estrutura organizacional voltada para a pesquisa, como uma universidade, instituto de pesquisa, ou mesmo uma empresa cujos empreendimentos dependem da gestão de dados.

TECNOLÓGICA - compreende um vasto conjunto de atividades, equipamentos, processos e expertises que possam viabilizar os requisitos tecnológicos operacionais necessários às ciberinfraestruturas de gestão de dados:

INFORMACIONAL – arcabouço conceitual e teórico materializado nas práticas da Ciência da Informação, Bublioteconomia, Arquivologia e Museologia aplicáveis a gestão de dados de pesquisa: tesauro, vocabulários, taxonomia, ontologias, bases de dados etc.

NORMAS E PADRÕES - Normas e padrões são formas consensuais de codificar o conhecimento que circula transversalmente por comunidades para assegurar uniformidade e similitude nos seus produtos e processos através do tempo e do espaço.

EXPERTISES - As instituições de pesquisa desenvolvem os mais diversos enfoques de gestão de dados. Isto pressupõe equipes de apoio compostas por diferentes profissionais. um requisito essencial é a necessidade de conhecimento das disciplinas e domínios nos quais os dados são coletados, processados e utilizados.

Papéis na gestão de dados de pesquisa

GESTOR– **administrador de C&T** que compreende a importância dos dados no âmbito institucionais, nacional e internacional, nessa direção apoia a definição de políticas, negocia recursos junto às agencias de fomento, implanta e-infraestruturas e adquire ferramentas e coleções de dados.

PESQUISADOR - **Autor/criador/coletor dos dados**; envolvido na pesquisa que produz os dados; o autor dos dados deve assegurar que os metadados, o registro dos dados, contexto e qualidade está em conformidade com os padrões da comunidade (NSC, 2005). Elabora junto com o bibliotecário/arquivista o PGD

BIBLIOTECÁRIO DE DADOS - Profissional da área de **biblioteconomia** com formação em gestão de repositórios de dados e de curadoria, indexação e catalogação de dados e conhecedor dos fluxos das pesquisas locais. Promove cursos e apóia a elaboração do PGD

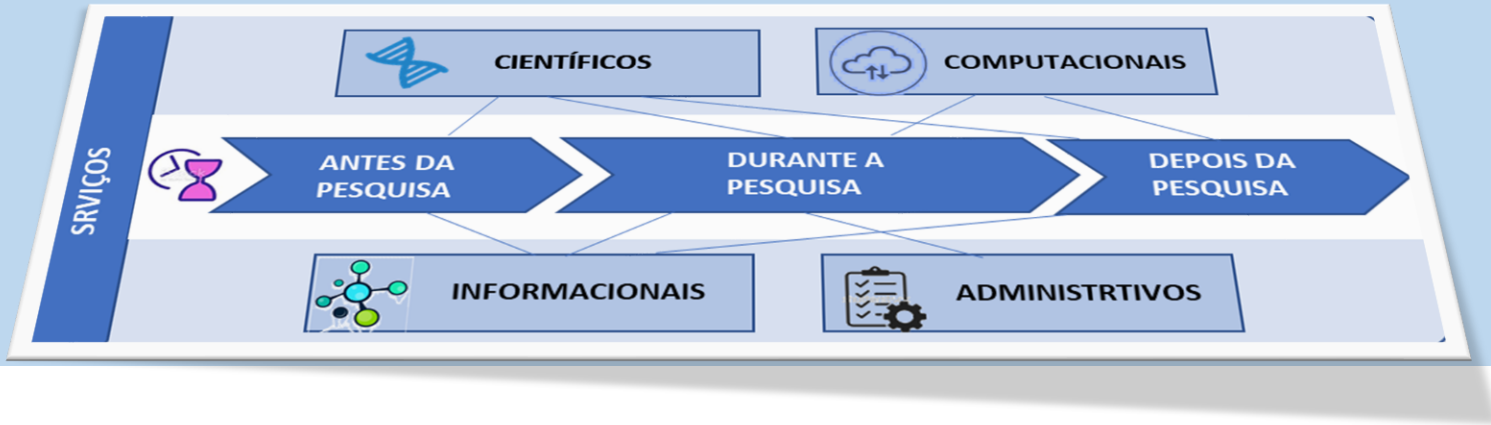
ARQUIVISTA DE DADOS – profissional de **arquivologia** responsável pelo arquivamento e preservação de longo prazo dos dados e garantia de autenticidade, integridade e confiabilidade

CIENTISTA DE DADOS – profissional das **áreas de computação** e/ou da área disciplinar que contribui no desenvolvimento de tecnologias de análise, manipulação, visualização, modelagem, algoritmos para as coleções de dados. Trabalha próximo aos pesquisadores

GERENTE DE DADOS – **tecnologista da informação** responsável pela manutenção e operação das bases de dados, segurança e armazenamento dos dados: backups, checagem de integridade, etc.

CURADOR DE DADOS – **pesquisador ou cientista de informação** com conhecimento disciplinar que adiciona valor aos dados por meio de documentação, integração, anotações, *mashup*, etc. Promove o compartilhamento e reuso, avalia para a preservação e cria serviços,

SERVIÇOS



CIENTÍFICOS

COMPUTACIONAIS

INFORMACIONAIS

ADMINISTRATIVOS

ANTES DA PESQUISA COMEÇAR

DURANTE A PESQUISA

PESQUISA FINALIZADA

SERVIÇOS

SERVIÇOS CIENTÍFICOS

Compreendem os serviços que estão **circunscritos ao ambiente científico**, como laboratórios, e executados por de **pesquisadores ou especialistas em gestão de dados com um alto grau de conhecimentos disciplinares**. São serviços relacionados à **preparação de dados para usos mais amplos e análises** e podem incluir atividades como, **descrição, documentação e contextualização das coleções de dados, avaliação, limpeza, normalização, organização dos arquivos, atribuição de nomes e, quando necessário, anonimização e outras estratégias para preservação da privacidade**. Mesmo considerando que esses serviços são protagonizados pelos próprios pesquisadores, eles precisam de considerável **suporte computacional** e, em muitos casos, serviços **informacionais**.

SERVIÇOS

SERVIÇOS CIENTÍFICOS

- ATRIBUIÇÃO DE METADADOS DISCIPLINARES
- AVALIAÇÃO DAS COLEÇÕES (APPRAISAL)
- LIMPEZA DOS DADOS
- ORGANIZAÇÃO DOS DADOS
- DOCUMENTAÇÃO DOS CÓDIGOS

- ANÁLISE DOS DADOS
- APRESENTAÇÃO E VISUALIZAÇÃO DOS DADOS
- TRANSFORMAÇÃO
- DOCUMENTAÇAO DO WORKFLOW
- DOCUMENTAÇÃO DO PROCESSAMENTO
- EMPACOTAMENTO DOS DADOS

SERVIÇOS

SERVIÇOS INFORMACIONAIS

Compreendem os serviços oferecidos pelos **profissionais de informação** no âmbito de organizações como bibliotecas científicas e centros de informação: i**dentificação** persistente de objetos de pesquisa e pesquisadores; desenvolvimento de **estruturas de representação como esquemas de metadados, taxonomia e ontologias; catalogação e indexação** de objetos de pesquisa; **publicação** de dados; **divulgação**; **letramento** de pesquisadores; **desenvolvimento de coleções de dados**; apoio à elaboração de **planos de gestão de dados**; **arquivamento de longo prazo para a preservação; *linking* e contextualização**.

SERVIÇOS

SERVIÇOS INFORMACIONAIS

COORDENAÇÃO DA CURADORIA

PORTAL DE GESTÃO DE DADOS

BALCAO DE REFERÊNCIA DE DADOS

APOIO A ELABORAÇÃO DO PLANO DE GESTÃO DE DADOS

APOIO A DESCOBERTA E ACESSO A COLEÇÃO DE DADOS

DESENVOLVIMENTO DE COLEÇÕES DE DADOS

DESENVOLVIMENTO DE METADADOS

CRIAÇÃO DE REFERÊNCIAS PADRONIZADAS

IDENTIFICAÇÃO DE DADOS E PESQUISADORES

CATALOGAÇÃO DAS COLEÇÕES DE DADOS

ARQUIVAMENTO DE LONGO PRAZO/PRESERVAÇÃO

PUBLICAÇÃO DE DADOS

CONTEXTUALIZAÇÃO/LINKING

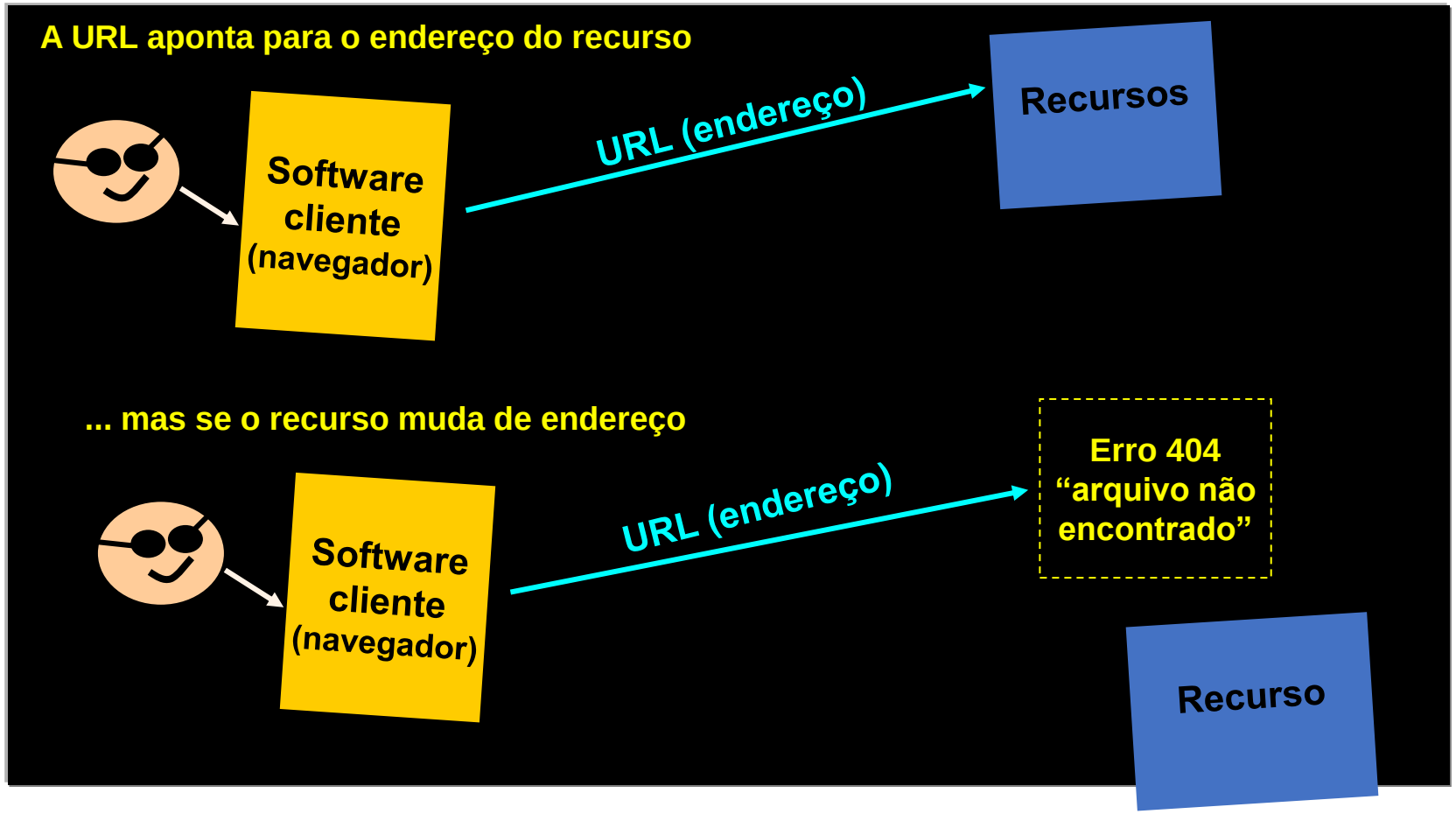
CAPACITAÇÃO

DIVULGAÇÃO

DISSEMINAÇÃO

IDENTIFICADORES PERSISTENTES

Os objetos digitais precisam ser identificados de forma persistente



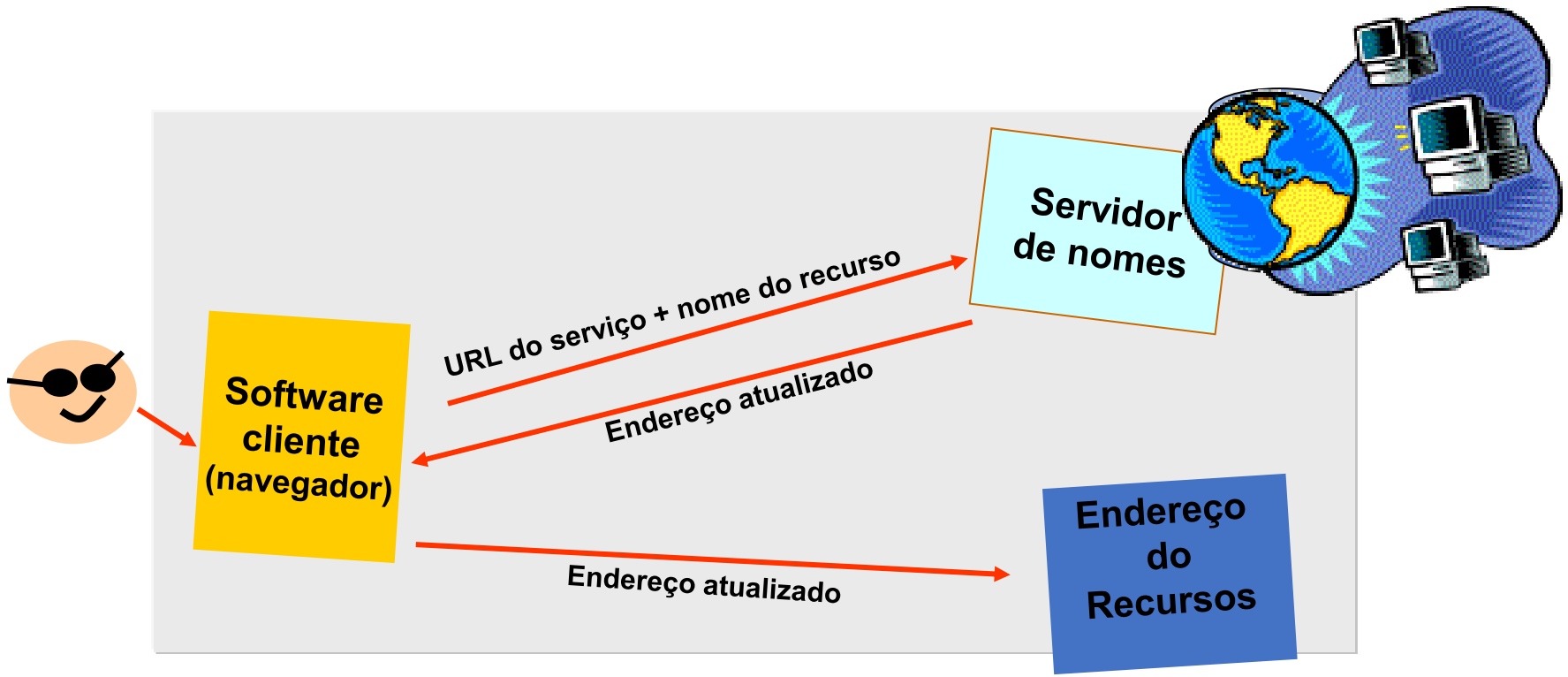
A URL é um endereço e não um identificador !

IDENTIFICADORES PERSISTENTES

Nomes devem ser:
UNICOS, GLOBAIS, PERSISTENTES, INDEPENDETES DE LOCALIZAÇÃO E TECNOLOGIA, PADRONIZADOS

Para isto funcionar é necessário estabelecer uma **infraestrutura administrativa** para decidir quem pode assinalar nomes que identificam univocamente os recursos digitais de forma persistente.

DOI, Handle System, PURL, URN ...

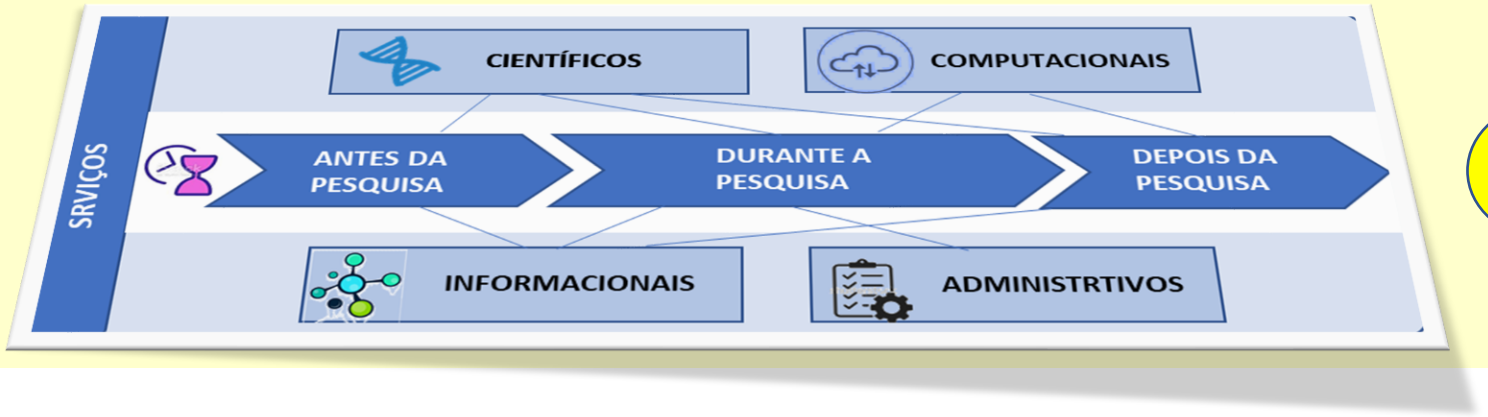


SERVIÇOS

SERVIÇOS E FERRAMENTAS COMPUTACIONAIS

Compreende a oferta **de ferramentas de *software* e recursos de computação** para apoiar o **processamento, análise e visualização dos dados** de pesquisa; **orientação de como os dados podem melhor ser estruturados e armazenados** e trabalhar, se necessário, junto aos pesquisadores na estruturação de bases de dados e marcação de texto etc.; os serviços podem incluir ainda **treinamento específico** para a equipe de pesquisadores nos recursos oferecidos e em situações mais avançadas, **oferecer processamento de alto desempenho**, bem como computação em grade.

SERVIÇOS



SERVIÇOS E FERRAMENTAS COMPUTACIONAIS

- **APOIO COMPUTACIONAL CIENTÍFICO** – disponibilidade de ferramentas de software e recursos de computação para apoiar o processamento, análise e visualização dos dados de pesquisa; o serviço pode, em situações mais avançadas, oferecer processamento de alto desempenho e computação em grade.
- **SISTEMA DE ARMAZENAMENTO** - tem como objetivo disponibilizar dispositivos de armazenamento para o amplo espectro de conjunto de dados gerados ou utilizadas pela instituição, numa escala que dê atenção aos usos corrente, mas que também antecipe os requisitos futuros das atividades de pesquisa das diversas equipes de pesquisadores; Os requisitos de desempenho dos sistemas de armazenamento podem variar em virtude dos níveis de utilização dos dados.
- **PROTEÇÃO DE DADOS SENSÍVEIS**
- **NORMALIZAÇÃO DE FORMATOS**
- **ESTRUTURAÇÃO DOS DADOS**
- **SERVIÇO DE BACKUP**
- **APOIO DE ELIMINAÇÃO DE DADOS**
- **CAPACITAÇÃO DOS PESQUISADORES**

SERVIÇOS

SERVIÇOS ADMINISTRATIVOS

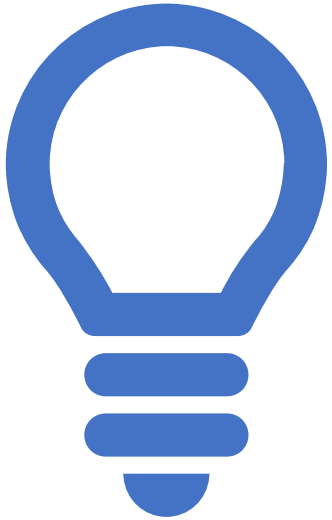
Compreende serviços de orientação sobre **custos**, **orçamento**, **aquisição de coleções de dados**, **conformidades ética e legal dos dados** – especialmente dados sensíveis – às normativas e regulamentos institucionais, nacionais e internacionais; **estatísticas** de uso e reuso dos dados; está categoria envolve também as questões de **propriedade intelectual**, **licenças e tempo de embargo**.

SERVIÇOS

SERVIÇOS ADMINISTRATIVOS

- **AQUISIÇÃO DE COLEÇÃO DE DADOS –** além de se utilizar de fontes externas, algumas instituições e suas bibliotecas estão adquirindo coleções individuais de dados motivado por demandas de seus pesquisadores. Este processo muitas vezes requerem uma extensa negociação em relação aos custos e à amplitude do acesso e dos termos de uso. As bibliotecas – especialmente as envolvidas com aquisição de recursos digitais – estão bem posicionadas para dar apoio nessa atividades.
- **CUSTO/ORÇAMENTO**
- **ESTATISTICA**
- **PROPRIEDADE INTELECTUAL**
- **LICENÇAS e TEMPO DE EMBARGO**
- **CONFORMIDADE ÉTICA E LEGAL**

O QUE O FAIR NÃO É



Esses princípios não têm a menor pretensão de impor qualquer tipo de padrão.

Eles não são somente RDF ou dados conectados e Web Semântica.

Eles podem ser igualmente utilizados por qualquer tipo de dado e serviço provenientes de qualquer disciplina.

E os dados FAIR não são a mesma coisa que dados abertos.

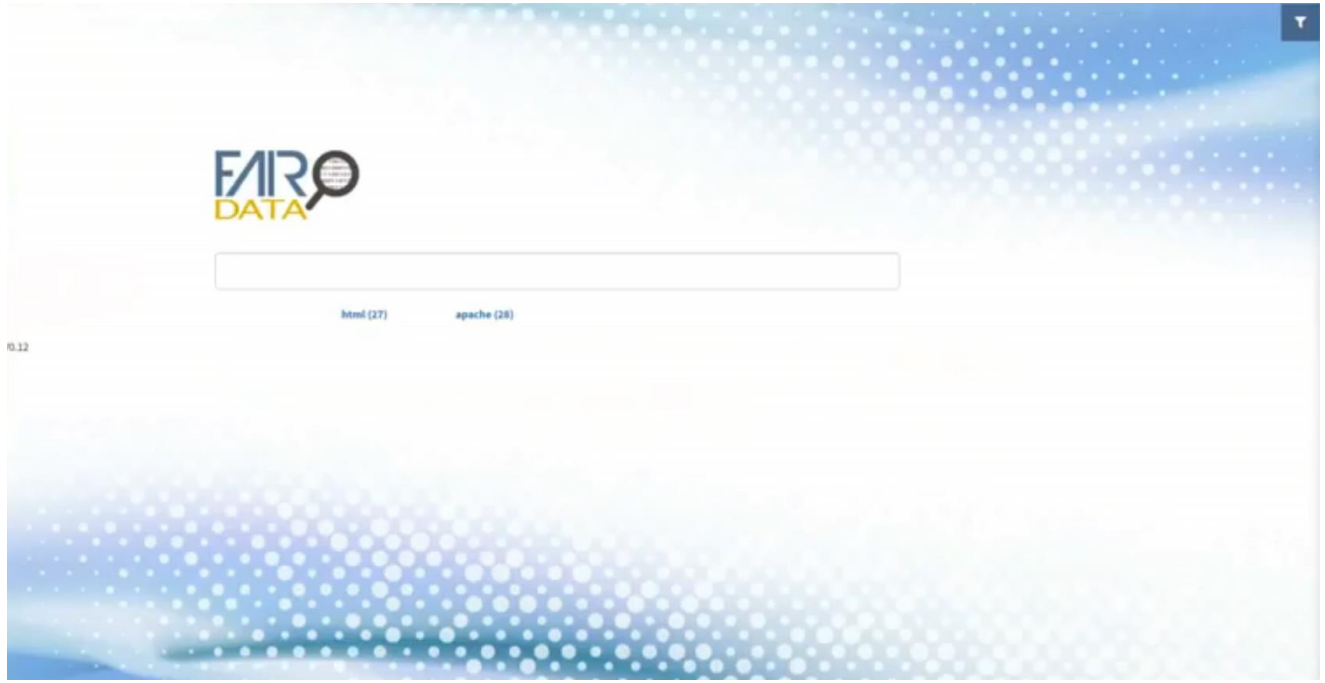
Como avaliar a encontrabilidade de seus dados?

Os seus dados estão associados a um **identificador persistênte**?

Existem metadados ricos descrevendo os seus dados?

Os metadados estão acessíveis online em uma ferramenta de busca? Ex: catálogo ou repositório de dados

O registro do metadado especifica o identificador persistente?



Como avaliar a acessibilidade de seu dado?

- Acessando o identificador persistente levará o usuário a dado ou a metadado associado?
- Os protocolos adotados seguem **padrões** reconhecidos? Ex: html
- Os procedimentos de acesso incluem autenticação e graus de autorização?
- Os metadados estão acessivies, mesmo quando os dados não estião disponíveis?

Como avaliar o nível de interoperabilidade de seus dados?

Os dados estão em formatos claramente compreendidos e de preferência abertos?

Os metadados seguem **normas** relevantes?

Os vocabulários controlados, palavras-chave, tesauros ou ontologias são utilizados sempre que possível?

Referências e links qualificados são fornecidos para uso por outros?

Como avaliar a nível de possibilidade de reuso dos seus dados?

- Os dados são precisos e bem descritos com muitos atributos relevantes?
- Os dados possuem uma licença de uso de dados clara e acessível?
- Está claro como, por que e por quem os dados foram criados e processados?
- Os dados e os metadados atendem a padrões e domínio relevante?

O Movimento GO-FAIR

É UM MOVIMENTO INTERNACIONAL QUE VISA DISSEMINAR E AUXILIAR A IMPLEMENTAÇÃO DOS PRINCÍPIOS FAIR.

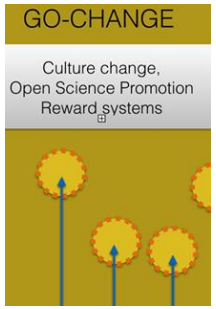
A Iniciativa

- GO (Global Open) FAIR é uma iniciativa que parte das comunidades de usuários com o objetivo de fazer com que dados fragmentados e desconectados sejam Encontráveis, Acessíveis, Interoperáveis e Reusáveis (FAIR na sigla em inglês) por máquinas e pessoas. 2016
- Enquanto iniciativa, o GO FAIR busca o desenvolvimento de um ambiente global compartilhado para a pesquisa e inovação baseada em dados.
- O centro da iniciativa é a federação de redes temáticas de excelência existentes que se compromete coletivamente aos Princípios FAIR em termos de estratégias de implementação desses princípios, o que incluem padrões, protocolos, políticas, diretrizes, boas práticas, etc.

- Mobiliza as redes existentes de excelência para convergir e 'falar a uma só voz' com os financiadores (públicos e privados) de pesquisa e inovação sobre questões anteriormente polêmicas, como a falta de investimento sustentável (a infraestrutura subjacente para 'ciência aberta)
- Reconhece que o caso para cada componente de um serviço individual (ex: BioPortal, uma única ontologia, FAIRSharing, ferramentas ISA etc.) existe um grau de dificuldade que é ampliado especialmente porque existe uma forte concorrência por fontes de financiamentos e que quando recebe-se o financiamento, o mesmo só dura alguns meses ou ano, não sendo possível manter o serviço instalado e funcionando além disso
- Reconhece também que tabelas de mapeamento, protocolos e outros padrões emergentes da comunidade não devem apenas **encontrar um 'lar'** (como por exemplo FAIRsharing), mas também devem ser **endossados coletivamente** e usados na prática por **comunidades** muito mais **coerentes** .
- O GO FAIR apoia o processo de coordenação dentro e entre as redes de implementação, treinamento e certificação a fim de **minimizar a reinvenção de componentes de infraestrutura redundantes**, incluindo coisas como tesauros e protocolos de ontologias genéricas ou específicas de domínio e outros elementos relacionados a padrões do IFDS

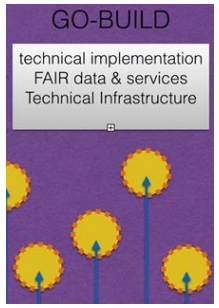
GO FAIR BRASIL – Quem somos?

- No Brasil, a primeira reunião aconteceu em setembro 2018, na cidade de São Paulo, no aniversário de 20 anos do Scielo, com lançamento do manifesto em dezembro do mesmo ano, em Brasília, nas dependencias do MCTI
- O GO FAIR Brasil tem a responsabilidade de difundir, apoiar e coordenar as atividades relacionadas à adoção da estratégia de implementação dos princípios FAIR definida pela iniciativa GO FAIR em todo o território brasileiro.
- O GO FAIR Brasil se compromete às seguintes atividades:
  - Apoiar e coordenar as Redes de Implementação, de acordo com seus objetivos específicos, a adotar as estratégias de implementação dos princípios FAIR aprovadas pela iniciativa GO FAIR;
  - Apoiar e coordenar Redes de Implementações que estejam dispostas a definir estratégias de implementação dos princípios FAIR nos casos em que não existam;
  - Sistematizar as diretrizes existentes e as criadas pelas Redes, garantindo que elas estejam condizentes com os princípios FAIR;
  - Desenvolver mecanismos de difusão das diretrizes definidas pelas Redes de Implementação;
  - Manter constante comunicação com o Escritório Internacional de Apoio e Coordenação GO FAIR



GO CHANGE – foca nas prioridades, políticas e incentiva a implementação do FAIR. Promove mudanças sócio-culturais que envolvem importantes atores de todos os níveis, visando tornar os princípios FAIR um padrão de trabalho na ciência.

GO TRAIN – trata da necessidade de definir currículos e programas de treinamento sobre o gerenciamento de dados. A meta é facilitar a capacitação de profissionais em gestão de dados.



GO BUILD – Trata da necessidade de criar infraestruturas para dados interoperáveis, criando padrões, protocolos e serviços compatíveis e possibilitando que o pesquisador deposite, acesse e analise dados científicos de todas as áreas.

Como estamos Organizados?

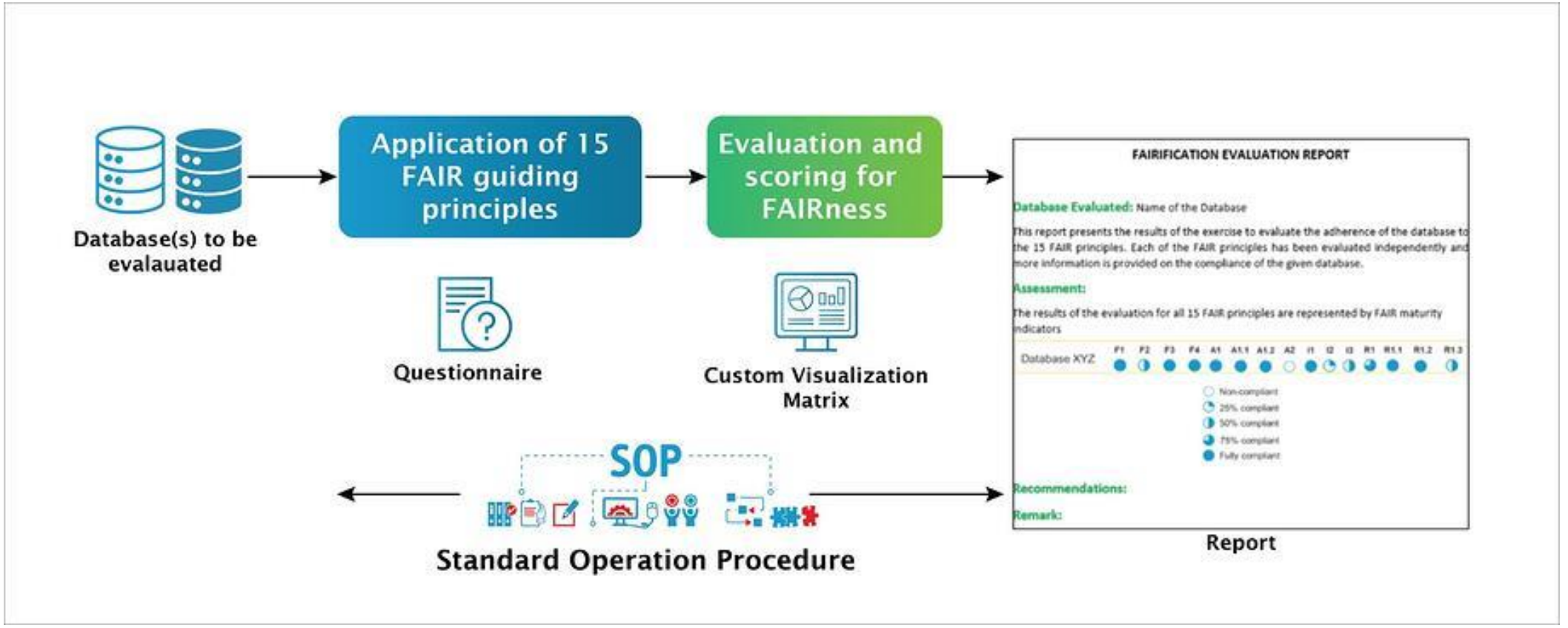
https://www.go-fair-brasil.org/

- Redes de Implementação
  - GOFAIR Brasil Saúde – Liderança FIOCRUZ
    - GO FAIR Brasil Enfermagem – Liderança UNIRIO
  - GO FAIR Brasil Humanidades – Liderança IBICT
  - GO FAIR Brasil Agro – Liderança EMBRAPA
  - GO FAIR Brasil Biodiversidade – Liderança Jardim Botânico
  - GO FAIR Brasil Ciências Nucleares – Liderança CNEN

Por que não indicamos um padrão

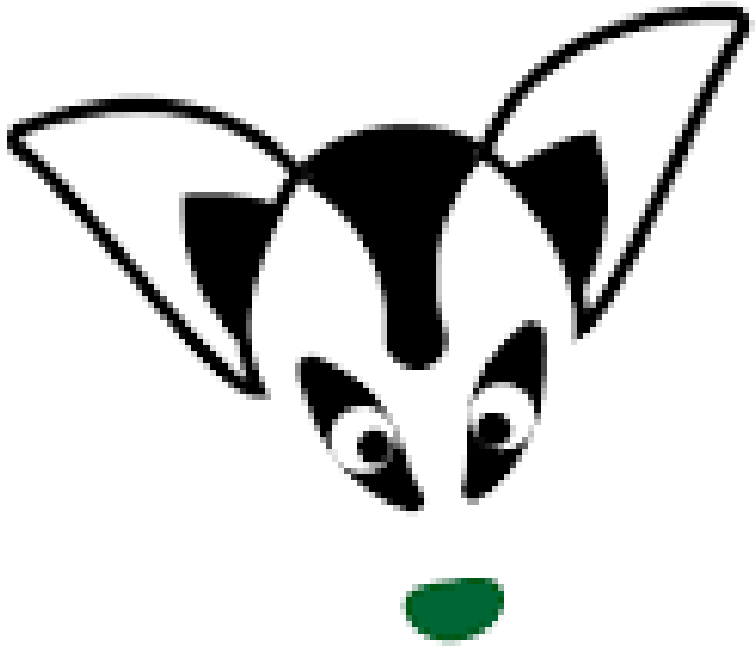
“Mas, como dissemos, aprendemos que, classicamente, os domínios operam em silos e que, mesmo dentro dos domínios, vários padrões, vocabulários, linguagens e abordagens continuarão a surgir. Isso não é apenas um incômodo e uma falta de coordenação e disciplina, é também uma **parte intrínseca do processo criativo** que deve ser apoiado a fim de promover nosso conhecimento e impulsionar a inovação. Isso significa que 'tabelas de mapeamento', 'bibliotecas para escolher', 'registros de padrões da comunidade', etc. continuarão a ser elementos cruciais da infraestrutura de suporte do IFDS”. (GO FAIR, 2020)

Como avaliar o nível de fairificação?



FAIR DATABR

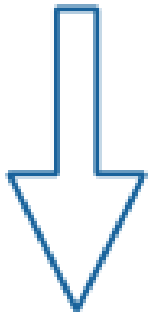
https://wrco.ufpb.br/fair/index.html



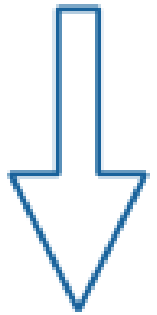
FairData BR

Create

Publish



Find



Annotate

DTL

DataFAIRport

À GUISA DE CONCLUSÃO

- Princípios FAIR devem ser aplicados não apenas à dados de pesquisa, mas à repositórios e outros objetos digitais
- O conceito de repositório deve ser visto como uma parte de uma Plataforma de gestão e serviços, oferecendo aos usuários outras formas de gerenciar seus dados e objetos digitais
- Tudo isso nos levará à uma Internet de Dados e Serviços que tornará o acesso à informaÇão e o processamento de conhecimento mais ágil e mais eficás


Live Ibict:
Infraestruturas de dados para pesquisas na emergência da COVID-19: desafios à abertura e ao compartilhamento
Sarita Albagli
Vanessa Jorge
Luis Sayão
Luana Sales
Quarta-feira, 20 de Outubro, às 16h
live.ibict.br
ibict
Instituto Brasileiro de Informação em Ciência e Tecnologia
UNIDADE DE PESQUISA DO MCTI
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES
PÁTRIA AMADA BRASIL
GOVERNO FEDERAL

SORTEIO

1 BOLSA 100%

1 BOLSA 50%

https://sorteador.com.br/