

ESCOLA DE INVERNO PPGC-UFF 2020 13 de agosto às 13h30

Curso de Metodologia de Pesquisa em Computação



Boas Práticas para o Compartilhamento de Dados de Pesquisa

O que é uma pesquisa F.A.I.R.?

O que é, para que serve, como se usa o DOI?

Como o compartilhamento de dados apoia a Ciência Aberta?

Troy Kohwalter 📆



Mais informações no site: http://ev-ppgc.ic.uff.br/2020-2/metodologia.html



Troy Costa Kohwalter troy@ic.uff.br





Compartilhamento de Dados de Pesquisa DOI Ética e Consentimento F.A.I.R. Ciência Aberta

AGENDA DESSA APRESENTAÇÃO





Dados de pesquisa Por que compartilhar? Como compartilhar?

COMPARTILHAMENTO DE DADOS DE PESQUISA





Dados de Pesquisa

Definição:

"Dados são fatos, observações ou experiências sobre as quais um argumento ou teoria é construído ou testado. Os dados podem ser numéricos, descritivos, auditivos ou visuais. Os dados podem ser brutos, abstraídos ou analisados, experimentais ou observacionais."

- Política de Dados de Pesquisa da UCL (*University College London*)



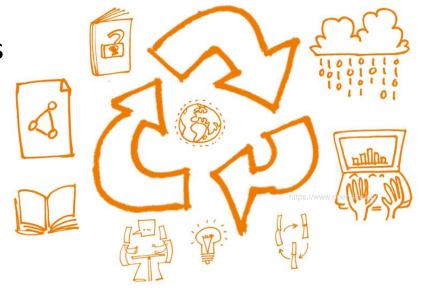
https://blog.opinionbox.com/





Dados de Pesquisa

- São as fontes ou materiais originais que você criou ou reuniu para conduzir seu projeto de pesquisa
- Podem ser digitais ou não digitais
- Resposta à sua Pergunta de Pesquisa é baseada na análise desses dados de pesquisa







Dados de Pesquisa

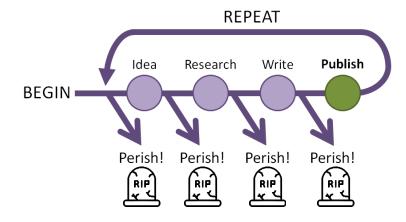
- Os dados de pesquisa são recursos valiosos!
- Geralmente exige muito tempo e dinheiro para serem produzidos
 - Pode levar diversos anos!
- Muitos dados têm um valor significativo além do uso para a pesquisa original
 - Colaborações
 - Surgimento de novas linhas de pesquisa



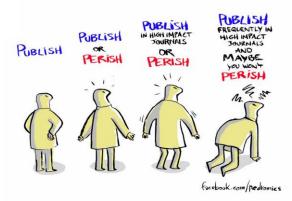




"Publish or Perish!"



THE EVOLUTION OF ACADEMIA







- Muitas instituições desejam compartilhar dados de pesquisa
 - Aumentar o impacto
 - Visibilidade de suas pesquisas
 - Disseminar o conhecimento
 - Estabelecer novas parcerias

"If we're all in a ship together, and the ship has some holes in it, and we're sort of bailing water out of it, and we have a great design for a bucket, then even if we're bailing out way better than everyone else, we should probably still share the bucket design."

-Elon Musk, 2014







Compartilhando os dados de pesquisa:

Incentiva a investigação científica e o debate

O que eu ganho com isso?

Promove inovação e novos usos potenciais dos dados

Leva a novas colaborações entre usuários e criadores de dados

Maximiza a transparência e responsabilidade

Incentiva a melhoria e validação de métodos de pesquisa

Permite o apuração dos resultados da pesquisa







Compartilhando os dados de pesquisa:

Reduz o custo de duplicação da coleta de dados

Aumenta o impacto e a visibilidade da pesquisa

Promove a pesquisa que criou os dados e seus resultados

Pode fornecer crédito ao pesquisador

Fornece recursos importantes para educação e treinamento

Aumenta o perfil acadêmico dos pesquisadores

O que eu ganho com isso?







Como compartilhar?

- Existem várias maneiras de compartilhar dados de pesquisa:
 - Colocar em um data center especializado
 - Colocar em um repositório institucional
 - Disponibilizar online por meio de um projeto ou site institucional
 - Disponibilizar informalmente entre pesquisadores
 - Escrever artigos científicos!
 - Submetê-los para uma revista ou conferencia













Conferências e Periódicos

- Exigem cada vez mais que os dados usados nas publicações sejam compartilhados ou disponibilizados em um banco de dados ou repositório acessível
 - Ajuda a tornar seus dados citáveis, rastreáveis e localizáveis!



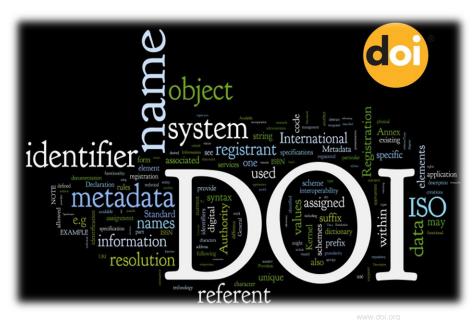




- Dados de pesquisa e as publicações baseadas nesses dados fazem parte da produção científica
 - Publicar é bastante importante para os pesquisadores!

























Digital Object Identifier

DOI









O que é DOI?

DOI = Digital Object Identifier (system) ®

- Organização governante: www.doi.org
 - International DOI Foundation ("IDF")
 - Organização sem fins lucrativos
 - Órgão de gestão para a federação de agências de registro (RAs)
 - Fornecem serviços e registro de DOI
 - É a autoridade de registro para o padrão ISO (ISO 26324) para o sistema DOI
- RAs = DOI Registration Agencies
 - Membros da IDF que oferecem o sistema DOI
 - Oferecerem um serviço baseado em DOI aos usuários
 - Para pessoas que desejam atribuir DOIs























O que é DOI?

- Sistema DOI é uma aplicação prática da ideia de repensar a Internet como gestão da informação
 - Não apenas como o movimento de pacotes de dados!

Tem alcance internacional!





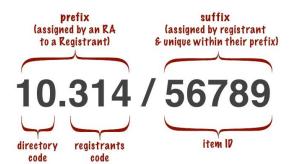


Para que serve o DOI?

Fornece um link resolvível, persistente e interoperável

Resolvível

Sintaxe de identificador padrão + mecanismo de resolução de rede



Exemplos da revista Elsevier:

https://doi.org/10.1016/j.is.2020.101610 https://doi.org/10.1016/j.entcom.2019.100318 https://doi.org/10.1016/j.entcom.2018.05.001 https://doi.org/10.1016/j.future.2018.03.004

Persistente

 Por meio de infraestrutura técnica (banco de dados de registro, suporte de proxy, etc) e infraestrutura social (obrigações das agências de registro)

Interoperável

Por meio de um modelo de dados (interoperabilidade semântica)





Para que serve o DOI?



Entertainment Computing

Volume 27, August 2018, Pages 110-127



e e interoperável

Understanding game sessions through

provenance

Troy Costa Kohwalter ^a 1 1 , Felipe Machado de Azeredo Figueira $^{b, 2}$ 1 , Juardo Assis de Lima Serdeiro $^{b, 2}$ 1 , Jose Ricardo da Silva Junior $^{b, 2}$ 2 , Leonardo Gresta Paulino Murta $^{a, 1}$ 1 , Esteban Valter Gonzalez Clua $^{a, 1}$ 1

Show more >

https://doi.org/10.1016/j.entcom.2018.05.001

Get rights and cente

canismo de resolução de rede

Exemplos da revista Elsevier:

https://doi.org/10.1016/j.is.2020.101610 https://doi.org/10.1016/j.entcom.2019.100318

https://doi.org/10.1016/j.entcom.2018.05.001

https://doi.org/10.1016/j.future.2018.03.004

Highlights

- Game analysis techniques normally do not consider cause-and-effect relationship.
- Concrete framework that uses provenance for representing cause-andeffect in games.
- Shows the analysis of a game session using the framework over two existing games.
- Provenance analysis in early stages of game development for game balancing.

anco de dados de registro, suporte de igações das agências de registro)

:eroperabilidade semântica)





Para que serve o DOI?

Fornece um link resolvível, persistente e interoperável

Resolvível

Sintaxe de identificador padrão + mecanismo de resolução de rede



Exemplos da revista Elsevier:

https://doi.org/10.1016/j.is.2020.101610 https://doi.org/10.1016/j.entcom.2019.100318 https://doi.org/10.1016/j.entcom.2018.05.001 https://doi.org/10.1016/j.future.2018.03.004

Persistente

 Por meio de infraestrutura técnica (banco de dados de registro, suporte de proxy, etc) e infraestrutura social (obrigações das agências de registro)

Interoperável

Por meio de um modelo de dados (interoperabilidade semântica)





Quem usa DOI?

- Os quatro principais tipos de usuários de DOI são:
 - 1. Editores
 - 2. Agências/afiliadas (mais de 3mil editoras)
 - 3. Bibliotecas
 - 4. Pesquisadores individuais









Benefícios do DOI

- CNPq usa o DOI como uma forma de certificação digital da produção científica
 - Informações fornecidas pelos pesquisadores em seus Currículos Lattes!
- Periódicos Científicos

Lattes

- Informa a localização de seus artigos
- Evita a perda de informação causada por alterações técnicas de sites, servidores, etc.
- Identificação segura e confiável!
 - Permite a articulação entre os serviços de provedores em diferentes níveis:
 - Título do artigo
 - Resumo
 - Referências
 - Citações





Benefícios do DOI

- Com DOI, podemos:
 - Olhar para catálogos de conteúdo disponível em toda a empresa
 - Pesquisar e procurar o conteúdo a ser utilizado
 - Ver onde o item existe dentro da organização
 - Recuperar o conteúdo







Dados Pessoais
Dados Confidenciais
Dados Pessoais Sensíveis
Declaração de Consentimento
Anonimização dos Dados



https://sebraeinteligenciasetorial.com.br/

ÉTICA E CONSENTIMENTO





- Dois fatores importantes para se compartilhar Dados de Pesquisa:
 - Consentimento
 - Confidencialidade









- Estratégias para lidar com a confidencialidade dependem da natureza da pesquisa
 - São essencialmente informadas pelas obrigações éticas e legais do pesquisador!

 O dever de confidencialidade para com os informantes pode ser explícito através de um termo assinado pelo participante





Tipos de dados:

Dados Pessoais

Dados Confidenciais

Dados Pessoais Sensíveis











Dados Pessoais

- São dados que se referem a um indivíduo
- Pode ser usado para identificar o indivíduo
 - Nome, sexo, endereço, telefone, CPF, etc.

Também inclui qualquer expressão de opinião sobre o

indivíduo







Dados Confidenciais

- São dados fornecidos em sigilo ou dados concordados em serem mantidos em sigilo
- Não são de domínio público







Dados Pessoais Confidenciais

- Raça
- Origem étnica
- Opinião política
- Crenças religiosas ou semelhantes
- Filiação a sindicatos
- Saúde ou condição física ou mental
- Etc.







Como conseguir o consentimento?

- Uma possibilidade é pelo TCLE
 - Termo de Consentimento Livre e Esclarecido
- É um documento que explica os procedimentos do experimento que deve ser assinado pelo participante
- Espera-se que os pesquisadores obtenham consentimento para que as pessoas participem da pesquisa e que possa usar as informações coletadas
- Sempre que possível, o consentimento também deve levar em consideração:
 - Qualquer uso futuro de dados
 - Compartilhamento
 - Preservação
 - Uso de longo prazo de dados de pesquisa

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE)



Condutor do Estudo: Pesquisadores Responsáveis Instituição:

Eventualmente realizamos estudos experimentais para caracterizar/avaliar uma determinada tecnologia de software. Estes estudos são conduzidos por [...]. Você foi previamente selecionado pelo seu perfil/conhecimento/experiência e está sendo convidado a participar desta pesquisa. Essa pesquisa consiste em avaliar [...].

1) Procedimentos

O estudo está sendo realizado com data e hora marcada com os participantes pré-selecionados. O estudo será executado de forma individual. O estudo consiste na [...]. Caso seja necessário, ao final do estudo será solicitado que você responda um questionário de avaliação sobre a elaboração do experimento.

2) Tratamento de possíveis riscos e desconfortos

Serão tomadas todas as providências durante a coleta de dados de forma a garantir a sua privacidade e seu anonimato.

3) Benefícios e Custos

Espera-se que, como resultado deste estudo, você possa aumentar seus conhecimentos, de maneira a contribuir para o aumento da qualidade das atividades com as quais você trabalhe ou possa vir a trabalhar. Este estudo também contribuirá com resultados importantes para a pesquisa de um modo geral. Você não terá nenhum gasto ou ônus com a sua participação no estudo e também não receberá qualquer espécie de reembolso ou gratificação devido à autorização do uso dos dados coletados nesse estudo.

4) Confidencialidade da Pesquisa

Toda informação coletada neste estudo é confidencial e seu nome não será identificado de modo algum. Quando os dados forem coletados, seu nome será removido dos mesmos e não será utilizado em nenhum momento durante a análise ou apresentação dos resultados.

5) Participação

Sua participação neste estudo é muito importante e voluntária, pois requer a sua aprovação para utilização dos dados coletados. Você tem o direito de não querer participar ou de sair deste estudo a qualquer momento, sem penalidades. Em caso de você decidir se retirar do estudo, favor notificar o pesquisador responsável. Você pode solicitar esclarecimento sobre o estudo a qualquer momento.

6) Declaração de Consentimento

Declaro que li e estou de acordo com as informações contidas neste documento e que toda linguagem técnica utilizada na descrição deste estudo de pesquisa foi explicada satisfatoriamente, recebendo respostas para todas as minhas dividas. Confirmo também que recebi uma cópia deste Termo (TCLE), compreendo que sou livre para não autorizar a utilização dos meus dados neste estudo em qualquer momento, sem qualquer penalidade. Declaro ter mais de 18 anos e conocrio de espontánea vontade em participar deste estudo.

Data://
Nome do Participante (letra de forma):
RG do Participante:
Assinatura:





- Os pesquisadores devem:
 - Informar aos participantes como os <u>dados da pesquisa serão armazenados</u>, preservados e usados a longo prazo
 - Questionários, vídeo, áudio, etc.
 - Informar aos participantes como a confidencialidade será mantida
 - Anonimizando dados
 - Obter consentimento informado, por escrito ou verbal, para compartilhamento de dados
 - Termo de consentimento
 - Explicito na gravação, caso seja verbal





Anonimização dos Dados

- Antes de usar os dados obtidos em pesquisa...
 - Verificar se precisa torná-los anônimo!
- Caso afirmativo:
 - Preservar a identidade dos participantes, organizações, empresas, etc.
 - Não ser possível identificar a pessoa a partir dos dados de pesquisa

De acordo com **London's Global University**, anonimato é o processo de remoção de identificadores pessoais, diretos e indiretos, que podem levar à identificação de um indivíduo.

FULLY IDENTIFIABLE DATA Maximum Utility Maximum Risk Acceptable Utility Acceptable Risk Minimum Utility Minimum Risk

DATA DEIDENTIFICATION





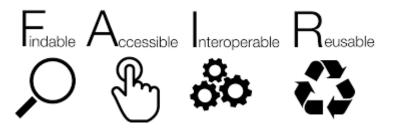
Anonimização dos Dados

- Anonimato pode ser necessário por razões éticas para proteger a identidade das pessoas
 - Também pode ser por razões legais para não divulgar dados pessoais
- Dados pessoais não devem ser divulgados a partir de informações de pesquisa!!!
 - Exceto quando o entrevistado tenha <u>dado consentimento explicito e</u> <u>específico para fazê-lo!</u>



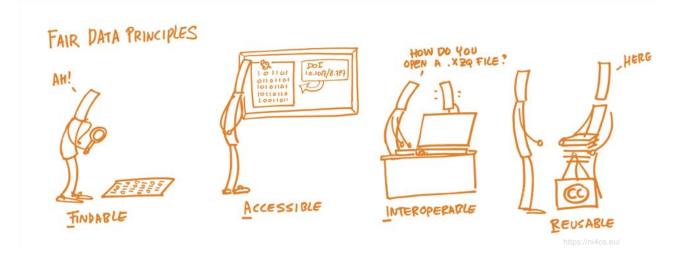






Findability
Accessibility
Interoperability
Reuse

F.A.I.R.

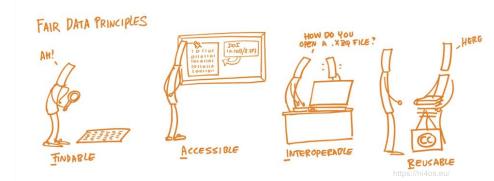






F.A.I.R.

- Findability, Accessibility, Interoperability, and Reuse
 - Criado em 2016
- Fornece diretrizes para melhorar:
 - Encontrabilidade
 - Acessibilidade
 - Interoperabilidade
 - Reutilização de ativos digitais



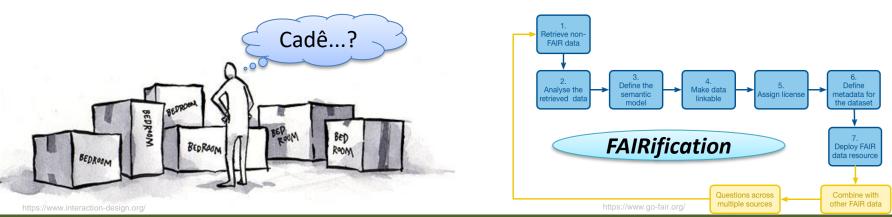
- Os princípios enfatizam a ação da máquina
 - A capacidade dos sistemas computacionais de <u>encontrar</u>, <u>acessar</u>, <u>interoperar</u> e <u>reutilizar</u> dados com nenhuma ou mínima intervenção humana
- Pois nós, humanos, dependemos cada vez mais do suporte computacional para lidar com os dados
 - Alto volume de dados existentes!
 - Complexidade crescente





Encontrabilidade

- O primeiro passo para (re)usar dados é encontrá-los!
- Metadados e dados devem ser fáceis de encontrar para humanos e computadores
- Metadados legíveis por máquina são essenciais para a descoberta automática
- Componente essencial do processo de FAIRification







Encontrabilidade

- F1: (Meta) dados são atribuídos a um identificador globalmente único e persistente
 - Ex.: Orcid, DOI
- **F2:** Os dados são descritos com metadados ricos (*definidos no R1*)
 - Informações descritivas sobre o contexto, qualidade e condição ou características dos dados
- F3: Metadados incluem claramente e explicitamente o identificador dos dados que descrevem
 - Metadados e o conjunto de dados que eles descrevem geralmente s\(\tilde{a}\) arquivos separados...
- F4: (Meta) dados são registrados ou indexados em um recurso pesquisável
 - Ex.: Google conseguir indexar as páginas para aparecer nas buscas





Acessibilidade

- Não basta apenas encontrar os dados
 - Precisa saber como eles podem ser acessados!
- Possivelmente inclui também a autenticação e autorização







Acessibilidade

- A1: (Meta) dados são recuperáveis por seu identificador usando um protocolo de comunicação padronizado
 - A recuperação de dados deve ser mediada sem ferramentas especializadas ou proprietárias ou métodos de comunicação
 - A maioria dos produtores de dados usa http(s) ou ftp
- A1.1: O protocolo é aberto, gratuito e universalmente implementável
 - Este critério impacta na escolha do repositório onde irá compartilhar os dados
- A1.2: O protocolo permite um procedimento de autenticação e autorização, quando necessário
 - Muitas vezes faz sentido solicitar aos usuários que criem uma conta de usuário para um repositório para definir direitos específicos do usuário
- A2: Os metadados são acessíveis, mesmo quando os dados não estão mais disponíveis
 - Datasets tendem a degradar ou desaparecer com o tempo
 - Relacionado aos problema de registro e indexação descritos em F4
 - Útil para rastrear pessoas, instituições ou publicações associadas à pesquisa original





Interoperabilidade

 Dados geralmente precisam ser integrados a outros dados

 Dados as vezes também precisam interoperar com aplicativos ou fluxos de trabalho para análise, armazenamento e processamento







Interoperabilidade

- I1: Os (meta) dados usam uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento
 - Humanos devem ser capazes de trocar e interpretar os dados uns dos outros
 - Dados também devem ser legíveis para máquinas, sem a necessidade de algoritmos, tradutores ou mapeamentos especializados ou ad hoc
 - Ex.: JSON, RDF
- I2: (Meta) dados usam vocabulários que seguem os princípios FAIR
 - Vocabulário usado precisa ser documentado e resolvido usando identificadores globais únicos e persistentes
 - Documentação precisa ser facilmente encontrada e acessível por qualquer pessoa que use o conjunto de dados
 - Ex.: FAIR Data Point
- **I3:** (Meta) dados incluem referências qualificadas a outros (meta) dados
 - Especificar se um conjunto de dados se baseia em outro conjunto de dados
 - Se conjuntos de dados adicionais são necessários para completar os dados
 - Se informações complementares são armazenadas em um conjunto de dados diferente





Reuso

 O objetivo final do FAIR é <u>otimizar a reutilização</u> <u>de dados</u>

- Metadados e dados devem ser bem descritos!
 - Assim eles podem ser replicados e/ou combinados em diferentes configurações





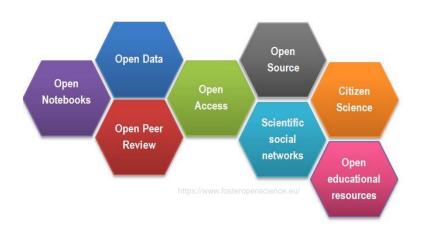




Reuso

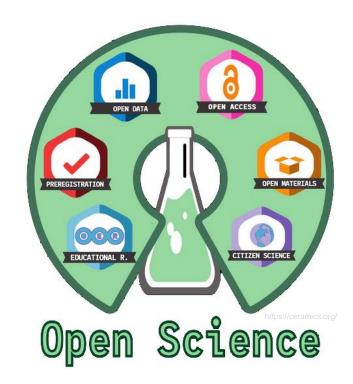
- R1: Meta (dados) são ricamente descritos com uma pluralidade de atributos precisos e relevantes
 - Está relacionado ao princípio F2, mas se concentra na capacidade de um usuário decidir se os dados são realmente úteis em um determinado contexto
 - O autor dos metadados deve ser o mais generoso possível, incluindo informações que podem parecer irrelevantes
 - Ex.: Descrever o escopo de seus dados, mencionar quaisquer particularidades ou limitações, especificar a data de geração/coleta dos dados
- R1.1: (Meta) dados são liberados com uma licença de uso de dados clara e acessível
 - Quais direitos de uso você atribui aos seus dados?
 - As condições sob as quais os dados podem ser usados devem ser claras para máquinas e humanos
 - Ex.: Licença MIT
- R1.2: (Meta) dados são associados à proveniência detalhada
 - Quem o gerou ou coletou? Como foi processado? Já foi publicado antes? Ele contém dados de outra pessoa que você pode ter transformado ou completado?
 - Incluir uma descrição do fluxo de trabalho que levou aos seus dados
- R1.3: (Meta) dados atendem aos padrões da comunidade relevantes ao domínio
 - Seguir os padrões estabelecidos pela comunidade para arquivamento e compartilhamento
 - Se divergir, justificar no metadado





Ciência aberta x Ciência Fechada Como seguir na direção de ciência aberta?

CIÊNCIA ABERTA









Qual caminho seguir?

Ciência aberta



Ciência fechada





Qual caminho seguir?

Ciência Aberta

- Disponibilizar conhecimento
- Universidades
- Artigo cientifico
- Mestrado e doutorado



Image courtesy of Robin Champieux

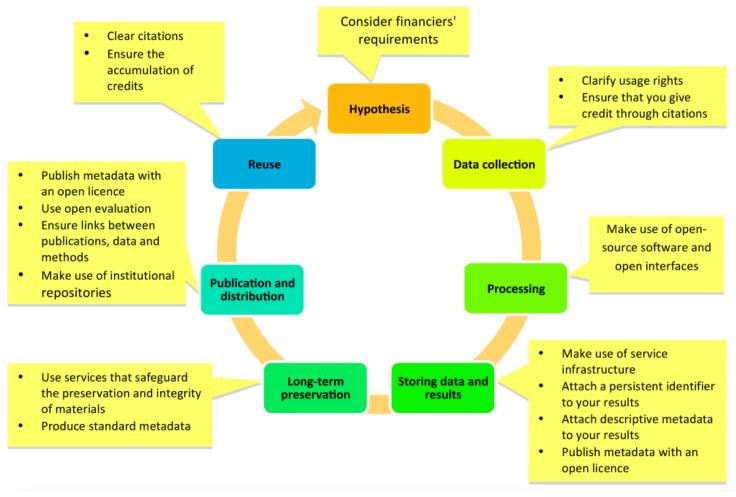
Ciência Fechada

- Pesquisa fechada
- Empresa privada
 - Investimento privado
- Patentes





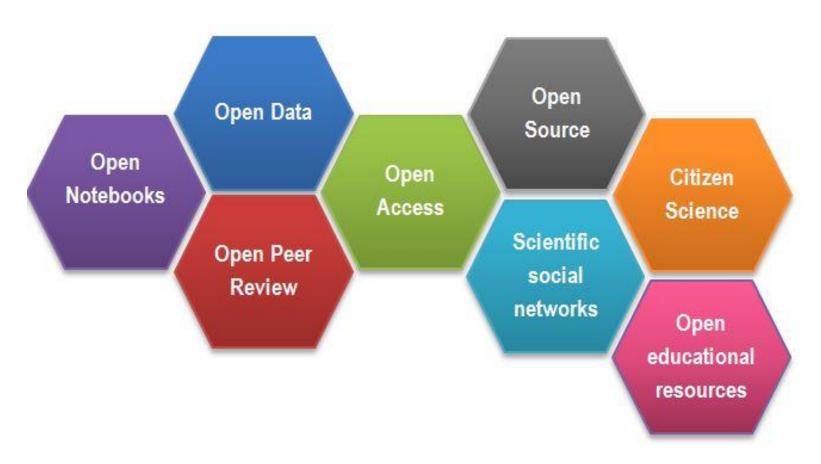
Ciência aberta...



https://www.fosteropenscience.eu/content/what-open-science-introduction



Ciência aberta...



https://www.fosteropenscience.eu/content/what-open-science-introduction





Como seguir na direção de ciência aberta?

- Adotar um serviço de hospedagem de código
 - Ex. GitHub
- Escolher uma licença apropriada
 - Ex. MIT
- Disponibilizar todos os materiais
 - Ex. código, dados, protocolos, exemplos, etc.
- Documentar adequadamente o projeto
- Facilitar o processo de uso/instalação
- Acolher potenciais usuários
- Publicar!





Licença MIT

Copyright <YEAR> <COPYRIGHT HOLDER>

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.





Como seguir na direção de ciência aberta?

- Adotar um serviço de hospedagem de código
 - Ex. GitHub
- Escolher uma licença apropriada
 - Ex. MIT
- Disponibilizar todos os materiais
 - Ex. código, dados, protocolos, exemplos, etc.
- Documentar adequadamente o projeto
- Facilitar o processo de uso/instalação
- Acolher potenciais usuários
- Publicar!





Por exemplo...







Principais referências bibliográficas

- Eynden V., et al., Managing and Sharing Data, 3ª edição, 2011
- Ferreira, Elisabete, et al., Digital Object Identifier (DOI): o que é, para que serve, como se usa?. AtoZ: novas práticas em informação e conhecimento 4.1 (2015): 5-9.
- London's Global University: https://www.ucl.ac.uk/
- Go FAIR: https://www.go-fair.org/



ESCOLA DE INVERNO PPGC-UFF 2020 13 de agosto às 13h30

Curso de Metodologia de Pesquisa em Computação



Boas Práticas para o Compartilhamento de Dados de Pesquisa

O que é uma pesquisa F.A.I.R.?

O que é, para que serve, como se usa o DOI?

Como o compartilhamento de dados apoia a Ciência Aberta?

Troy Kohwalter 📆



Mais informações no site: http://ev-ppgc.ic.uff.br/2020-2/metodologia.html



Troy Costa Kohwalter troy@ic.uff.br