Aqui está a tradução do PDF fornecido:

**Os Princípios Orientadores FAIR para gestão e administração de dados científicos**

**Mark D. Wilkinson et al.**

Recebido: 10 de dezembro de 2015 1

Aceito: 12 de fevereiro de 2016 2

Publicado: 15 de março de 2016 3

Há uma necessidade urgente de melhorar a infraestrutura que apoia a reutilização de dados acadêmicos. 4 Um conjunto diversificado de partes interessadas — representando a academia, a indústria, as agências de fomento e os editores acadêmicos — uniu-se para projetar e endossar conjuntamente um conjunto conciso e mensurável de princípios que chamamos de Princípios de Dados FAIR. 5 A intenção é que estes possam atuar como uma diretriz para aqueles que desejam aumentar a reutilização de seus acervos de dados. 6 Diferentemente de iniciativas de pares que se concentram no pesquisador humano, os Princípios FAIR dão ênfase específica em aprimorar a capacidade das máquinas de encontrar e usar automaticamente os dados, além de apoiar sua reutilização por indivíduos. 7 Este Comentário é a primeira publicação formal dos Princípios FAIR e inclui a lógica por trás deles e algumas implementações exemplares na comunidade. 8

**Apoiando a descoberta através de uma boa gestão de dados**

A boa gestão de dados não é um fim em si mesma, mas sim o principal canal que leva à descoberta de conhecimento e à inovação, e à subsequente integração e reutilização de dados e conhecimento pela comunidade após o processo de publicação de dados. 9 Infelizmente, o ecossistema digital existente em torno da publicação de dados acadêmicos nos impede de extrair o máximo benefício de nossos investimentos em pesquisa (por exemplo, ref. 1). 10 Parcialmente em resposta a isso, os financiadores da ciência, editores e agências governamentais estão começando a exigir planos de gestão e administração de dados para dados gerados em experimentos financiados publicamente. 11 Além da coleta, anotação e arquivamento adequados, a administração de dados inclui a noção de "cuidado a longo prazo" de ativos digitais valiosos, com o objetivo de que eles sejam descobertos e reutilizados para investigações posteriores, seja sozinhos ou em combinação com dados recém-gerados. 12 Os resultados de uma boa gestão e administração de dados, portanto, são publicações digitais de alta qualidade que facilitam e simplificam este processo contínuo de descoberta, avaliação e reutilização em estudos subsequentes. 13 O que constitui uma "boa gestão de dados" é, no entanto, largamente indefinido e geralmente deixado como uma decisão para o proprietário dos dados ou do repositório. 14 Portanto, trazer alguma clareza em torno dos objetivos e desiderata de uma boa gestão e administração de dados, e definir marcos simples para informar aqueles que publicam e/ou preservam dados acadêmicos, seria de grande utilidade. 15

Este artigo descreve quatro princípios fundamentais — Encontrabilidade (Findability), Acessibilidade (Accessibility), Interoperabilidade (Interoperability) e Reutilização (Reusability) — que servem para orientar os produtores e editores de dados enquanto navegam por esses obstáculos, ajudando assim a maximizar o valor agregado obtido pela publicação acadêmica digital formal e contemporânea. 16 É importante ressaltar que nossa intenção é que os princípios se apliquem não apenas a "dados" no sentido convencional, mas também aos algoritmos, ferramentas e fluxos de trabalho que levaram a esses dados. 17 Todos os objetos de pesquisa digital acadêmica — de dados a pipelines analíticos — se beneficiam da aplicação desses princípios, uma vez que todos os componentes do processo de pesquisa devem estar disponíveis para garantir transparência, reprodutibilidade e reutilização. 18 Existem inúmeras e diversas partes interessadas que podem se beneficiar da superação desses obstáculos: pesquisadores que desejam compartilhar, obter crédito e reutilizar os dados e interpretações uns dos outros; 19 editores de dados profissionais que oferecem seus serviços; desenvolvedores de software e ferramentas que fornecem serviços de análise e processamento de dados, como fluxos de trabalho reutilizáveis; 20 agências de fomento (privadas e públicas) cada vez mais preocupadas com a administração de dados a longo prazo; e uma comunidade de ciência de dados que minera, integra e analisa dados novos e existentes para avançar na descoberta. 21 Para facilitar a leitura deste manuscrito por essas diversas partes interessadas, fornecemos definições para abreviações comuns no Quadro 1. 22

**Quadro 1 | Termos e Abreviações**

* **BD2K** - Big Data 2 Knowledge, é uma iniciativa trans-NIH estabelecida para permitir a pesquisa biomédica como uma empresa de pesquisa digital, para facilitar a descoberta e apoiar novos conhecimentos e para maximizar o engajamento da comunidade. 23
* **DOI** - Digital Object Identifier; um código usado para identificar de forma permanente e estável objetos (geralmente digitais). 24 Os DOIs fornecem um mecanismo padrão para a recuperação de https://www.google.com/search?q=metadados sobre o objeto e, geralmente, um meio de acessar o próprio objeto de dados. 25
* **FAIR** - Findable, Accessible, Interoperable, Reusable (Encontrável, Acessível, Interoperável, Reutilizável). 26
* **FORCE11** - The Future of Research Communications and e-Scholarship; uma comunidade de acadêmicos, bibliotecários, arquivistas, editores e financiadores de pesquisa que surgiu organicamente para ajudar a facilitar a mudança em direção a uma melhor criação e compartilhamento de conhecimento, iniciada em 2011. 27
* **Interoperabilidade** - a capacidade de dados ou ferramentas de recursos não cooperantes de se integrarem ou trabalharem juntos com o mínimo de esforço.
* **JDDCP** - Joint Declaration of Data Citation Principles, Reconhecendo os dados como um produto de pesquisa de primeira classe e para apoiar boas práticas de pesquisa em torno da reutilização de dados, a JDDCP propõe um conjunto de princípios orientadores para a citação de dados na literatura acadêmica, em outro conjunto de dados ou em qualquer outro objeto de pesquisa. 28
* **RDF** - Resource Description Framework, uma estrutura globalmente aceita para representação de dados e conhecimento que se destina a ser lida e interpretada por máquinas. 29

Os humanos, no entanto, não são as únicas partes interessadas críticas no meio dos dados científicos. 30 Problemas semelhantes são encontrados pelas aplicações e agentes computacionais que encarregamos de realizar a recuperação e análise de dados em nosso nome. 31 Essas "partes interessadas computacionais" são cada vez mais relevantes e exigem tanta ou mais atenção à medida que sua importância cresce. 32 Um dos grandes desafios da ciência intensiva em dados, portanto, é melhorar a descoberta de conhecimento, auxiliando tanto os humanos quanto seus agentes computacionais na descoberta, acesso, integração e análise de dados científicos apropriados para a tarefa e outros objetos digitais acadêmicos. 33

Para certos tipos de objetos digitais importantes, existem repositórios especializados, bem curados e profundamente integrados, como o Genbank, o Worldwide Protein Data Bank (wwPDB) e o UniProt nas ciências da vida; 34 o Space Physics Data Facility (SPDF; <http://spdf.gsfc.nasa.gov/>) e o Set of Identifications, Measurements and Bibliography for Astronomical Data (SIMBAD) nas ciências espaciais. 35 Esses recursos centrais, fundamentais e críticos, estão continuamente curando e capturando conjuntos de dados de referência de alto valor e ajustando-os para aprimorar a produção acadêmica, fornecer suporte para usuários humanos e mecânicos e fornecer ferramentas extensas para acessar seu conteúdo de maneiras ricas e dinâmicas. 36 No entanto, nem todos os conjuntos de dados ou mesmo tipos de dados podem ser capturados por, ou submetidos a, esses repositórios. 37 Muitos conjuntos de dados importantes que emergem da ciência de bancada tradicional e de baixo rendimento não se encaixam nos modelos de dados desses repositórios de propósito específico, mas esses conjuntos de dados não são menos importantes no que diz respeito à pesquisa integrativa, reprodutibilidade e reutilização em geral. 38

Aparentemente em resposta a isso, vemos o surgimento de numerosos repositórios de dados de propósito geral, em escalas que vão do institucional (por exemplo, uma única universidade) a repositórios abertos de escopo global, como Dataverse, FigShare ([http://figshare.com](http://figshare.com/)), Dryad, Mendeley Data (<https://data.mendeley.com/>), Zenodo (<http://zenodo.org/>), DataHub ([http://datahub.io](http://datahub.io/)), DANS (<http://www.dans.knaw.nl/>) e EUDat. 39 Tais repositórios aceitam uma ampla gama de tipos de dados em uma grande variedade de formatos, geralmente não tentam integrar ou harmonizar os dados depositados e impõem poucas restrições (ou requisitos) aos descritores do depósito de dados. 40 O ecossistema de dados resultante, portanto, parece estar se afastando da centralização, tornando-se mais diverso e menos integrado, exacerbando assim o problema de descoberta e reutilização para as partes interessadas humanas e computacionais. 41

Um exemplo específico desses obstáculos pode ser imaginado no domínio da regulação e análise da expressão gênica. 42 Suponha que um pesquisador tenha gerado um conjunto de dados de sítios de poliadenilação diferencialmente selecionados em um organismo patogênico não modelo cultivado sob uma variedade de condições ambientais que estimulam seu estado patogênico. 43 O pesquisador está interessado em comparar os genes alternativamente poliadenilados neste conjunto de dados local com outros exemplos de poliadenilação alternativa e os níveis de expressão desses genes tanto neste organismo quanto em organismos modelo relacionados durante o processo de infecção. 44 Dado que não há um arquivo de propósito específico para dados de poliadenilação diferencial e nenhum banco de dados de organismo modelo para este patógeno, por onde o pesquisador começa? 45 Consideraremos a abordagem atual para este problema a partir de uma variedade de perspectivas de descoberta e integração de dados. 46 Se os conjuntos de dados desejados existissem, onde eles poderiam ter sido publicados e como alguém começaria a procurá-los, usando quais ferramentas de busca? 47 A busca desejada precisaria filtrar com base em espécies específicas, tecidos específicos, tipos específicos de dados (Poly-A, microarray, NGS), condições específicas (infecção), e os genes específicos — essa informação ('https://www.google.com/search?q=metadados') é capturada pelos repositórios e, em caso afirmativo, em que formatos está, é pesquisável e como? 48 Uma vez que os dados são descobertos, eles podem ser baixados? 49 Em que formato(s)? 50 Esse formato pode ser facilmente integrado com dados privados internos (o conjunto de dados local de sítios de poliadenilação alternativa), bem como com outras publicações de dados de terceiros e com os repositórios de dados de genes/proteínas centrais da comunidade? 51 Essa integração pode ser feita automaticamente para economizar tempo e evitar erros de copiar e colar? 52 O pesquisador tem permissão para usar os dados desses pesquisadores de terceiros, sob quais condições de licença, e quem deve ser citado se um ponto de dados for reutilizado? 53

Questões como essas destacam algumas das barreiras à descoberta e reutilização de dados, não apenas para humanos, mas ainda mais para máquinas; 54 no entanto, são precisamente esses tipos de análises profundamente e amplamente integrativas que constituem a maior parte da e-Science contemporânea. 55 A razão pela qual muitas vezes precisamos de várias semanas (ou meses) de esforço técnico especializado para reunir os dados necessários para responder a tais questões de pesquisa não é a falta de tecnologia apropriada; 56 a razão é que não damos aos nossos valiosos objetos digitais a atenção cuidadosa que eles merecem quando os criamos e preservamos. 57

Superar essas barreiras, portanto, requer que todas as partes interessadas — incluindo pesquisadores, repositórios de propósito específico e de propósito geral — evoluam para enfrentar os desafios emergentes descritos acima. 58 O objetivo é que os objetos digitais acadêmicos de todos os tipos se tornem "cidadãos de primeira classe" no ecossistema de publicação científica, onde a qualidade da publicação — e, mais importante, o impacto da publicação — é uma função de sua capacidade de ser encontrada, reutilizada e citada de forma precisa e apropriada ao longo do tempo, por todas as partes interessadas, tanto humanas quanto mecânicas. 59

Com esse objetivo em mente, um workshop foi realizado em Leiden, Holanda, em 2014, chamado 'Jointly Designing a Data Fairport'. 60 Este workshop reuniu um amplo grupo de partes interessadas acadêmicas e privadas, todas com interesse em superar os obstáculos de descoberta e reutilização de dados. 61 Das deliberações no workshop, surgiu a noção de que, através da definição e do amplo apoio a um conjunto mínimo de princípios e práticas orientadoras acordados pela comunidade, todas as partes interessadas poderiam descobrir, acessar, integrar e reutilizar adequadamente e citar adequadamente as vastas quantidades de informação geradas pela ciência contemporânea intensiva em dados. 62 A reunião foi concluída com a formulação de um rascunho de um conjunto de princípios fundamentais que foram subsequentemente elaborados em maior detalhe - a saber, que todos os objetos de pesquisa devem ser Encontráveis, Acessíveis, Interoperáveis e Reutilizáveis (FAIR), tanto para máquinas quanto para pessoas. 63 Estes são agora referidos como os Princípios Orientadores FAIR. 64 Posteriormente, um grupo de trabalho FAIR dedicado, estabelecido por vários membros da comunidade FORCE11, aprimorou e melhorou os Princípios. 65 Os resultados desses esforços são relatados aqui. 66

**A importância das máquinas em ambientes de pesquisa ricos em dados**

A ênfase colocada na aplicação do FAIR tanto para atividades humanas quanto para atividades orientadas por máquinas é um foco específico dos Princípios Orientadores FAIR que os distingue de muitas iniciativas de pares (discutidas na seção subsequente). 67 Humanos e máquinas frequentemente enfrentam barreiras distintas ao tentar encontrar e processar dados na Web. 68 Os humanos têm um senso intuitivo de "semântica" (o significado ou a intenção de um objeto digital) porque somos capazes de identificar e interpretar uma ampla variedade de pistas contextuais, sejam elas na forma de pistas estruturais/visuais/icônicas no layout de uma página da Web ou no conteúdo de notas narrativas. 69 Como tal, somos menos propensos a cometer erros na seleção de dados apropriados ou outros objetos digitais, embora os humanos enfrentem dificuldades semelhantes se faltarem https://www.google.com/search?q=metadados contextuais suficientes. 70 A principal limitação dos humanos, no entanto, é que não somos capazes de operar no escopo, escala e velocidade necessários pela escala dos dados científicos contemporâneos e pela complexidade da e-Science. 71 É por essa razão que os humanos dependem cada vez mais de agentes computacionais para realizar tarefas de descoberta e integração em seu nome. 72

Isso requer que as máquinas sejam capazes de agir de forma autônoma e apropriada quando confrontadas com a ampla gama de tipos, formatos e mecanismos/protocolos de acesso que encontrarão durante sua exploração autoguiada do ecossistema global de dados. 73 Também exige que as máquinas mantenham um registro primoroso de proveniência para que os dados que estão coletando possam ser citados de forma precisa e adequada. 74 Auxiliar esses agentes, portanto, é uma consideração crítica para todos os participantes no processo de gerenciamento e administração de dados, desde pesquisadores e produtores de dados até os anfitriões de repositórios de dados. 75

Ao longo deste artigo, usamos a frase "acionável por máquina" para indicar um contínuo de possíveis estados nos quais um objeto digital fornece informações cada vez mais detalhadas a um explorador de dados computacional que atua de forma autônoma. 76 Essa informação permite ao agente, em um grau que depende da quantidade de detalhes fornecidos, ter a capacidade, ao se deparar com um objeto digital nunca antes encontrado, de: a) identificar o tipo de objeto (com respeito tanto à estrutura quanto à intenção), b) determinar se é útil no contexto da tarefa atual do agente, interrogando https://www.google.com/search?q=metadados e/ou elementos de dados, c) determinar se é utilizável, com respeito à licença, consentimento ou outras restrições de acessibilidade ou uso, e d) tomar a ação apropriada, da mesma forma que um humano faria. 77 Por exemplo, uma máquina pode ser capaz de determinar o tipo de dados de um objeto digital descoberto, mas não ser capaz de analisá-lo por estar em um formato desconhecido; 78 ou pode ser capaz de processar os dados contidos, mas não ser capaz de determinar os requisitos de licenciamento relacionados à recuperação e/ou uso desses dados. 79 O estado ideal — onde as máquinas "entendem" totalmente e podem operar de forma autônoma e correta em um objeto digital — pode raramente ser alcançado. 80 No entanto, os princípios FAIR fornecem "passos ao longo de um caminho" em direção à acionabilidade por máquina; 81 adotar, no todo ou em parte, os princípios FAIR, conduz o recurso ao longo do contínuo em direção a este estado ideal. 82

Além disso, a ideia de ser acionável por máquina se aplica em dois contextos — primeiro, ao se referir aos https://www.google.com/search?q=metadados contextuais que cercam um objeto digital ('o que é?'), e segundo, ao se referir ao conteúdo do próprio objeto digital ('como eu o processo/integro?'). 83 Qualquer um, ou ambos, podem ser acionáveis por máquina, e cada um forma seu próprio contínuo de acionabilidade. 84 Finalmente, desejamos fazer uma distinção entre dados que são acionáveis por máquina como resultado de um investimento específico em software que suporta esse tipo de dados, por exemplo, parsers personalizados que entendem arquivos wwPDB de ciências da vida ou arquivos SPASE (Space Physics Archive Search and Extract) de ciências espaciais, e dados que são acionáveis por máquina exclusivamente através da utilização de tecnologias abertas de propósito geral. 85 Para reiterar o ponto anterior — a acionabilidade por máquina final ocorre quando uma máquina pode tomar uma decisão útil sobre dados que nunca encontrou antes. 86

Essa distinção é importante ao considerar (a) o ambiente de dados em rápida evolução e crescimento, com novas tecnologias e novos tipos de dados mais complexos sendo continuamente desenvolvidos, e (b) o crescimento de repositórios de propósito geral, onde os tipos de dados que um agente provavelmente encontrará são imprevisíveis. 87 Criar parsers personalizados, em todas as linguagens de computador, para todos os tipos de dados e todas as ferramentas analíticas que requerem esses tipos de dados, não é uma atividade sustentável. 88 Como tal, o foco em auxiliar as máquinas em sua descoberta e exploração de dados através da aplicação de tecnologias e padrões de interoperabilidade mais generalizados no nível de dados/repositório, torna-se uma primeira prioridade para uma boa administração de dados. 89

**Quadro 2 | Os Princípios Orientadores FAIR**

**Para ser Encontrável (Findable):**

* F1. (https://www.google.com/search?q=meta)dados recebem um identificador globalmente único e persistente.
* F2. dados são descritos com https://www.google.com/search?q=metadados ricos (definidos por R1 abaixo). 90
* F3. https://www.google.com/search?q=metadados incluem de forma clara e explícita o identificador dos dados que descrevem. 91
* F4. [cite\_start](https://www.google.com/search?q=meta)dados92 são registrados ou indexados em um recurso pesquisável. 93

**Para ser Acessível (Accessible):**

* A1. cite\_startdados são recuperáveis por seu identificador usando um protocolo de comunicação padronizado. 94
  + A1.1. o protocolo é aberto, gratuito e universalmente implementável.
  + A1.2. o protocolo permite um procedimento de autenticação e autorização, quando necessário.
* A2. https://www.google.com/search?q=metadados são acessíveis, mesmo quando os dados não estão mais disponíveis. 95

**Para ser Interoperável (Interoperable):**

* I1. (https://www.google.com/search?q=meta)dados usam uma linguagem formal, acessível, compartilhada e amplamente aplicável para representação do conhecimento.
* I2. [cite\_start](https://www.google.com/search?q=meta)dados usam vocabulários que seguem os princípios FAIR. 96
* I3. (https://www.google.com/search?q=meta)dados incluem referências qualificadas a outros (https://www.google.com/search?q=meta)dados.

**Para ser Reutilizável (Reusable):**

* R1. https://www.google.com/search?q=meta(dados) são ricamente descritos com uma pluralidade de atributos precisos e relevantes. 97
  + R1.1. [cite\_start](https://www.google.com/search?q=meta)dados são divulgados com uma licença de uso de dados clara e acessível. 98
  + R1.2. (https://www.google.com/search?q=meta)dados estão associados a uma proveniência detalhada.
  + R1.3. [cite\_start](https://www.google.com/search?q=meta)dados atendem aos padrões comunitários relevantes ao domínio. 99

**Os Princípios Orientadores FAIR em detalhes**

Representantes dos grupos de partes interessadas interessadas, discutidos acima, se uniram em torno de quatro desiderata centrais — os Princípios Orientadores FAIR e sua elaboração limitada, que foram refinados (Quadro 2) a partir do rascunho original da reunião, disponível em (<https://www.force11.org/node/6062>). 100 Um documento separado que aborda dinamicamente a discussão da comunidade relacionada a esclarecimentos e explicações dos princípios, e diretrizes detalhadas e exemplos de implementações FAIR, está sendo construído atualmente (<http://datafairport.org/fair-principles-living-document-menu>). 101

Os Princípios Orientadores FAIR descrevem considerações distintas para ambientes de publicação de dados contemporâneos com respeito ao suporte tanto ao depósito manual quanto automatizado, exploração, compartilhamento e reutilização. 102 Embora tenha havido uma série de publicações recentes, muitas vezes focadas em domínios, que defendem melhorias específicas nas práticas relacionadas ao gerenciamento e arquivamento de dados 1,11,12, o FAIR se diferencia por descrever princípios concisos, independentes de domínio e de alto nível que podem ser aplicados a uma ampla gama de produtos acadêmicos. 103 Ao longo dos Princípios, usamos a frase '(https://www.google.com/search?q=meta)dados' nos casos em que o Princípio deve ser aplicado tanto aos https://www.google.com/search?q=metadados quanto aos dados. 104

Os elementos dos Princípios FAIR são relacionados, mas independentes e separáveis. 105 Os Princípios definem características que os recursos de dados, ferramentas, vocabulários e infraestruturas contemporâneas devem exibir para auxiliar na descoberta e reutilização por terceiros. 106 Ao definir minimamente cada princípio orientador, a barreira de entrada para produtores, editores e administradores de dados que desejam tornar seus acervos de dados FAIR é mantida propositalmente o mais baixa possível. 107 Os Princípios podem ser seguidos em qualquer combinação e de forma incremental, à medida que os ambientes de publicação dos provedores de dados evoluem para graus crescentes de 'FAIRness'. 108 Além disso, a modularidade dos Princípios e sua distinção entre dados e https://www.google.com/search?q=metadados apoiam explicitamente uma ampla gama de circunstâncias especiais. 109 Um desses exemplos são os dados altamente sensíveis ou pessoalmente identificáveis, onde a publicação de https://www.google.com/search?q=metadados ricos para facilitar a descoberta, incluindo regras claras sobre o processo de acesso aos dados, fornece um alto grau de 'FAIRness' mesmo na ausência da publicação FAIR dos próprios dados. 110 Um segundo exemplo envolve a publicação de objetos de pesquisa que não são dados. 111 Fluxos de trabalho analíticos, por exemplo, são um componente crítico do ecossistema acadêmico, e sua publicação formal é necessária para alcançar tanto a transparência quanto a reprodutibilidade científica. 112 Os princípios FAIR podem ser igualmente aplicados a esses ativos que não são dados, que precisam ser identificados, descritos, descobertos e reutilizados da mesma forma que os dados. 113

Esforços exemplares específicos que fornecem níveis variados de FAIRness são detalhados posteriormente neste documento. 114 Questões adicionais, no entanto, permanecem a ser abordadas. Primeiro, quando os vocabulários endossados pela comunidade ou outros padrões de (https://www.google.com/search?q=meta)dados não incluem os atributos necessários para alcançar uma anotação rica, existem duas soluções possíveis: publicar uma extensão de um vocabulário existente e intimamente relacionado ou, no caso extremo, criar e publicar explicitamente um novo recurso de vocabulário, seguindo os princípios FAIR ('12'). 115 Em segundo lugar, para identificar explicitamente o padrão escolhido quando mais de um vocabulário ou outro padrão de (https://www.google.com/search?q=meta)dados está disponível, e dado que, por exemplo, nas ciências da vida existem mais de 600 padrões de conteúdo, o registro BioSharing (<https://biosharing.org/>) pode ser útil, pois descreve os padrões em detalhes, incluindo versões, quando aplicável. 116

**Os Princípios precedem a implementação**

Esses Princípios Orientadores FAIR de alto nível precedem as escolhas de implementação e não sugerem nenhuma tecnologia, padrão ou solução de implementação específica; 117 além disso, os Princípios não são, em si, um padrão ou uma especificação. 118 Eles atuam como um guia para editores e administradores de dados para auxiliá-los na avaliação se suas escolhas de implementação particulares estão tornando seus artefatos de pesquisa digital Encontráveis, Acessíveis, Interoperáveis e Reutilizáveis. 119 Prevemos que esses princípios de alto nível permitirão uma ampla gama de comportamentos integrativos e exploratórios, baseados em uma ampla gama de escolhas e implementações de tecnologia. 120 De fato, muitos repositórios já estão implementando vários aspectos do FAIR usando uma variedade de escolhas de tecnologia e vários exemplos são detalhados na próxima seção; 121 exemplos incluem o próprio Scientific Data e como artigos de dados narrativos são ancorados a https://www.google.com/search?q=metadados estruturados progressivamente FAIR. 122

**Exemplos de FAIRness e o valor agregado resultante**

* **Dataverse**: O Dataverse é um software de repositório de dados de código aberto instalado em dezenas de instituições em todo o mundo para apoiar repositórios comunitários públicos ou repositórios de dados de pesquisa institucionais. 123 O Harvard Dataverse, com mais de 60.000 conjuntos de dados, é o maior dos repositórios Dataverse atuais e está aberto a todos os pesquisadores de todos os campos de pesquisa. 124 O Dataverse gera uma citação formal para cada depósito, seguindo o padrão definido por Altman e King¹³. 125 O Dataverse torna público o Identificador de Objeto Digital (DOI), ou outros identificadores persistentes (Handles), quando o conjunto de dados é publicado ('F'). 126 Isso resolve para uma página de destino, fornecendo acesso a https://www.google.com/search?q=metadados, arquivos de dados, termos do conjunto de dados, isenções ou licenças e informações de versão, tudo isso indexado e pesquisável ('F', 'A' e 'R'). 127 Os depósitos incluem https://www.google.com/search?q=metadados, arquivos de dados e quaisquer arquivos complementares (como documentação ou código) necessários para entender os dados e a análise ('R'). 128 Os https://www.google.com/search?q=metadados são sempre públicos, mesmo que os dados sejam restritos ou removidos por questões de privacidade ('F', 'A'). 129 Esses https://www.google.com/search?q=metadados são oferecidos em três níveis, apoiando extensivamente os princípios FAIR 'I' e 'R': 1) https://www.google.com/search?q=metadados de citação de dados, que mapeiam para o esquema DataCite ou Dublin Core Terms, 2) https://www.google.com/search?q=metadados específicos de domínio, que quando possível mapeiam para padrões de https://www.google.com/search?q=metadados usados em um domínio científico, e 3) https://www.google.com/search?q=metadados em nível de arquivo, que podem ser profundos e extensos para arquivos de dados tabulares (incluindo https://www.google.com/search?q=metadados em nível de coluna). 130 Finalmente, o Dataverse fornece interfaces públicas acessíveis por máquina para pesquisar os dados, acessar os https://www.google.com/search?q=metadados e baixar os arquivos de dados, usando um token para conceder acesso quando os arquivos de dados são restritos ('A'). 131
* **FAIRDOM** (<http://fair-dom.org/about>): integra as plataformas SEEK¹⁴ e openBIS¹⁵ para produzir uma instalação de gerenciamento de dados e modelos FAIR para Biologia de Sistemas. 132 Ativos de pesquisa individuais (ou agregados de dados e modelos) são identificados com URLs HTTP únicos e persistentes, que podem ser registrados com DOIs para publicação ('F'). 133 Os ativos podem ser acessados pela Web em uma variedade de formatos apropriados para indivíduos e/ou seus computadores (RDF, XML) ('I'). 134 Os ativos de pesquisa são anotados com https://www.google.com/search?q=metadados ricos, usando padrões, formatos e ontologias da comunidade ('I'). 135 Os https://www.google.com/search?q=metadados são armazenados como RDF para permitir a interoperabilidade e os ativos podem ser baixados para reutilização ('R'). 136
* **ISA¹⁶**: é uma estrutura de rastreamento de https://www.google.com/search?q=metadados orientada pela comunidade para facilitar a coleta, curadoria, gerenciamento e reutilização de conjuntos de dados de ciências da vida em conformidade com os padrões. 137 A ISA fornece https://www.google.com/search?q=metadados estruturados progressivamente FAIR para os artigos de Descritor de Dados do Nature Scientific Data e muitos artigos de dados do GigaScience, e sustenta o banco de dados EBI MetaboLights, entre outros recursos de dados. 138 No cerne está um modelo ISA de propósito geral e extensível, originalmente disponível apenas como uma representação tabular, mas posteriormente aprimorado como uma representação baseada em RDF¹⁷ e serializações JSON para habilitar o 'I' e o 'R', tornando-se 'FAIR' quando publicado como dados vinculados ([http://elixir-uk.org/node-events/201cisa-as-a-fair-research-object201d-hack-the-spec-event-1](https://www.google.com/search?q=http://elixir-uk.org/node-events/201cisa-as-a-fair-research-object201d-hack-the-spec-event-1)) e complementando outros objetos de pesquisa¹⁸. 139
* **Open PHACTS¹⁹**: O Open PHACTS é uma plataforma de integração de dados para informações pertencentes à descoberta de medicamentos. 140 O acesso à plataforma é mediado por uma interface acessível por máquina²⁰ que fornece múltiplas representações que são legíveis tanto por humanos (HTML) quanto por máquinas (RDF, JSON, XML, CSV, etc), fornecendo a faceta 'A' da FAIRness. 141141141 A interface permite que várias URLs sejam usadas para acessar informações sobre uma entidade específica por meio de um serviço de mapeamento ('F' e 'A'). 142 Assim, um usuário pode fornecer uma URL do ChEMBL para recuperar informações originadas, por exemplo, do Chemspider ou do DrugBank. 143 Cada chamada fornece uma URL canônica em sua resposta ('A' e 'I'). 144 Todas as fontes de dados usadas são descritas usando descrições de conjunto de dados padronizadas, seguindo o padrão global VoID, com rica proveniência ('R' e 'I'). 145 Todos os recursos da interface são descritos usando RDF seguindo a especificação da API de Dados Vinculados ('A'). 146 Finalmente, a maioria dos conjuntos de dados é descrita usando ontologias acordadas pela comunidade ('I'). 147
* **wwPDB⁴,²¹**: O wwPDB é um arquivo de dados de propósito específico, intensivamente curado, que hospeda informações sobre estruturas 3D determinadas experimentalmente de proteínas e ácidos nucleicos. 148 Todas as entradas do wwPDB são hospedadas de forma estável em um servidor FTP ('A') e representadas em formatos legíveis por máquina (texto e XML); 149 este último é acionável por máquina usando os https://www.google.com/search?q=metadados fornecidos pelo wwPDB em conformidade com o Macromolecular Information Framework (mmCIF²²), um padrão de dados da União Internacional de Cristalografia (IUCr) ('F', 'I' para humanos, 'F', 'I' para máquinas cientes da IUCr). 150 Os https://www.google.com/search?q=metadados do wwPDB contêm referências cruzadas a identificadores comuns, como PubMed e NCBI Taxonomy, e seus https://www.google.com/search?q=metadados do wwPDB são descritos em dicionários de dados e documentos de esquema ([http://mmcif.wwpdb.org](http://mmcif.wwpdb.org/) e [http://pdbml.wwpdb.org](http://pdbml.wwpdb.org/)) que estão em conformidade com o padrão de dados da IUCr para os domínios da química e biologia estrutural ('R'). 151 Uma variedade de ferramentas de software está disponível para interpretar tanto os dados quanto os https://www.google.com/search?q=metadados do wwPDB ('I', 'R' para humanos, 'I', 'R' para máquinas com este software). 152 Cada entrada é representada por um DOI ('F', 'A' para humanos e máquinas). 153 O DOI resolve para um arquivo compactado que requer software especial para interrogatório/interpretação adicional. Outros pontos de acesso do wwPDB²³⁻²⁵ fornecem acesso aos registros do wwPDB por meio de URLs que provavelmente serão estáveis a longo prazo ('F'), e todos os dados e https://www.google.com/search?q=metadados são pesquisáveis por meio de um ou mais dos sites afiliados ao wwPDB ('F'). 154
* **UniProt²⁶**: O UniProt é um recurso abrangente para dados de sequência e anotação de proteínas. 155 Todas as entradas são identificadas exclusivamente por uma URL estável, que fornece acesso ao registro em uma variedade de formatos, incluindo uma página da web, texto simples e RDF ('F' e 'A'). 156 O registro contém https://www.google.com/search?q=metadados ricos ('F') que são legíveis tanto por humanos (HTML) quanto por máquinas (texto e RDF), onde a resposta formatada em RDF utiliza vocabulários e ontologias compartilhados, como UniProt Core, FALDO e ECO ('I'). 157 Com interligação com mais de 150 bancos de dados diferentes, cada registro do UniProt possui extensos links para, por exemplo, o PubMed, permitindo uma rica citação. 158 Esses links são acionáveis por máquina na representação RDF ('R'). 159 Finalmente, na representação RDF, a Ontologia Central do UniProt classifica explicitamente todos os registros, não deixando ambiguidade — nem para humanos nem para máquinas — sobre o que os dados representam ('R'), permitindo a recuperação totalmente automatizada de registros e informações de referência cruzada. 160

**Quadro 3 | Iniciativas comunitárias/colaborativas emergentes com o FAIR como foco ou atividade principal**

* **bioCADDIE (**[**https://biocaddie.org**](https://biocaddie.org/)**):** O consórcio NIH BD2K biomedical and healthCAre Data Discovery Index Ecosystem (bioCADDIE) trabalha para desenvolver um protótipo de Índice de Descoberta de Dados (DDI), que está previsto para ser tão transformador e impactante para os dados quanto o PubMed é para a literatura biomédica. 161 O DDI foca em encontrar ('F') e acessar ('A') os conjuntos de dados armazenados em diferentes fontes, e trabalha progressivamente para identificar https://www.google.com/search?q=metadados relevantes ('I') e mapeá-los para padrões da comunidade ('R'), vinculando-se ao BioSharing. 162
* **CEDAR:** O Center for Expanded Data Annotation and Retrieval (CEDAR) é um centro de excelência financiado pelo NIH BD2K para desenvolver ferramentas e tecnologias que reduzem o ônus de criar e aprimorar https://www.google.com/search?q=metadados que atendam aos padrões baseados na comunidade. O CEDAR permitirá a criação de modelos de https://www.google.com/search?q=metadados que implementam padrões baseados na comunidade para https://www.google.com/search?q=metadados experimentais, do BioSharing ([https://biosharing.org](https://biosharing.org/)), e que serão identificáveis e recuperáveis de forma única com URIs HTTP, e anotados com vocabulários e ontologias extraídos do BioPortal ([http://bioportal.bioontology.org](http://bioportal.bioontology.org/)) ('F', 'A', 'I', 'R'). 163 Esses modelos guiarão os usuários a criar https://www.google.com/search?q=metadados ricos com identificadores HTTP únicos e estáveis ('F') que podem ser recuperados usando HTTP ('A') e acessíveis em uma variedade de formatos (JSON-LD, TURTLE, RDF/XML, CSV, etc) ('I'). 164 Esses https://www.google.com/search?q=metadados usarão padrões da comunidade, conforme definido pelo modelo, e incluirão proveniência e uso de dados ('R'). 165
* **Esses dois projetos, entre outros, fornecem ferramentas e/ou oportunidades de colaboração para aqueles que desejam melhorar a FAIRness de seus dados.** 166

Além de, e em apoio a, comunidades e recursos que já estão buscando objetivos FAIR, o Grupo de Implementação de Citação de Dados da Force11 publicou recomendações técnicas específicas sobre como implementar muitos dos princípios, com um foco particular em identificadores e sua resolução, persistência e acessibilidade de https://www.google.com/search?q=metadados, especialmente relacionadas à citação. 167 Além disso, o grupo 'Skunkworks' que surgiu do Workshop Lorentz tem criado softwares que suportam infraestruturas que são, de ponta a ponta, compatíveis com os princípios FAIR, e podem ser implementados sobre repositórios existentes. 168 Esses módulos de código têm um foco particular na publicação e pesquisabilidade de https://www.google.com/search?q=metadados, compatibilidade em casos de considerações estritas de privacidade e no problema extremamente difícil da interoperabilidade de dados e https://www.google.com/search?q=metadados (manuscrito em preparação). 169 Finalmente, existem vários projetos emergentes, alguns listados no Quadro 3, para os quais o FAIR é um objetivo chave. 170 Esses projetos podem fornecer conselhos e orientações valiosos para aqueles que desejam se tornar mais FAIR. 171

**FAIRness é um pré-requisito para o gerenciamento e a administração adequados de dados**

As ideias dentro dos Princípios Orientadores FAIR refletem, combinam, baseiam-se e estendem trabalhos anteriores tanto dos parceiros da Concept Web Alliance (<https://conceptweblog.wordpress.com/>), que se concentraram na acionabilidade por máquina e na harmonização de estruturas e semântica de dados, quanto das organizações científicas e acadêmicas que desenvolveram a Declaração Conjunta de Princípios de Citação de Dados (JDDCP²⁹), que se concentrou em dados acadêmicos primários sendo tornados citáveis, detectáveis e disponíveis para reutilização, de modo a serem capazes de apoiar uma bolsa de estudos mais rigorosa. 172172172172 Uma tentativa de definir as semelhanças e sobreposições entre os Princípios FAIR e o JDDCP é fornecida em (<https://www.force11.org/node/6062>). 173

Os Princípios FAIR também são complementares ao 'Selo de Aprovação de Dados' (DSA) ([http://datasealofapproval.org/media/filer\_public/2013/09/27/guidelines\_2014-2015.pdf](https://www.google.com/search?q=http://datasealofapproval.org/media/filer_public/2013/09/27/guidelines_2014-2015.pdf)), na medida em que compartilham o objetivo geral de tornar os dados reutilizáveis por outros usuários além daqueles que os geraram originalmente. 174 Enquanto o DSA foca principalmente nas responsabilidades e conduta dos produtores de dados e repositórios, o FAIR foca principalmente nos próprios dados. 175 Claramente, a comunidade mais ampla de partes interessadas está se unindo em torno de um conjunto de visões comuns e interligadas que abrangem todas as facetas do ecossistema de publicação de dados acadêmicos. 176 O resultado final, quando implementado, será um gerenciamento e uma administração mais rigorosos desses valiosos recursos digitais, para o benefício de toda a comunidade acadêmica. 177

Como afirmado no início, uma boa gestão e administração de dados não é um fim em si mesma, mas sim uma pré-condição que apoia a descoberta e a inovação do conhecimento. 178 A e-Science contemporânea exige que os dados sejam Encontráveis, Acessíveis, Interoperáveis e Reutilizáveis a longo prazo, e esses objetivos estão se tornando rapidamente expectativas de agências e editores. 179 Demonstramos, portanto, que os Princípios de Dados FAIR fornecem um conjunto de marcos para produtores e editores de dados. 180 Eles orientam a implementação dos níveis mais básicos de boas práticas de Gerenciamento e Administração de Dados, ajudando assim os pesquisadores a aderirem às expectativas e requisitos de suas agências de fomento. 181 Convocamos todos os produtores e editores de dados a examinar e implementar esses princípios, e a participar ativamente da iniciativa FAIR, juntando-se ao grupo de trabalho da Force11. 182 Trabalhando juntos em direção a objetivos compartilhados e comuns, os valiosos dados produzidos por nossa comunidade alcançarão gradualmente os objetivos críticos da FAIRness. 183