Tabela 5. Resumo dos links que não estão mais disponíveis, discriminados por tipo de link.

Tipo de Link	Links que Falham ao Resolver	Total de Links desse Tipo	Porcentagem de Links com Falha desse Tipo
URL	79	1342	5.9%
DOI	13	744	1.7%
FTP	5	21	23.8%
ZIP	7	16	43.8%
PDF	7	12	58.3%
DOCX	2	9	22.2%
GZ	3	7	42.9%
XLSX	0	5	0.0%
URL mal formatada	1	3	33.3%
TXT	0	3	0.0%
DOC	0	2	0.0%
IMG	0	1	0.0%
ZIPR	1	1	100.0%

https://doi.org/10.1371/journal.pone.0304781.t005

arquivos provavelmente não estavam mais disponíveis (58,3% de perda). A taxa de perda de PDF é complicada pelo fato de que muitos desses links apontam para informações suplementares em vez de conjuntos de dados compartilhados; isso significa que as informações do link foram registradas erroneamente no CaltechAUTHORS (arquivos de informações suplementares deveriam ter sido carregados no repositório em vez de registrados como links de dados suplementares). Como cerca de um quarto dos artigos de autoria da Caltech contêm informações suplementares, uma quantidade significativa das quais é formatada como PDF, a disponibilidade contínua de dados PDF na internet merece mais estudos para uma melhor quantificação.

Como a disponibilidade de dados evolui ao longo do tempo? (RQ2.2)

Ao examinar os links que estavam indisponíveis por ano de publicação do artigo (ver Fig 2), um padrão emerge onde links mais antigos eram mais propensos a estarem indisponíveis do que links mais novos. O número de links testados variou entre os anos de publicação – com menos links de dados suplementares testados para publicações mais antigas – portanto, apenas dados de 2014–2022 são plotados e ajustados na Fig 2, pois esses anos tiveram pelo menos 10 links para testar. Modelando a relação entre a disponibilidade dos conjuntos de dados e a idade do artigo usando uma regressão de Poisson, verificou-se que as chances de os dados estarem disponíveis foram reduzidas em 2,6% (razão de chances $= 0.974$ [.950-.998, IC 95%], $p < 0.05$) para cada ano após a publicação do artigo.

Existem sites específicos onde os dados são mais propensos a desaparecer? (RQ2.3)

As Tabelas 6 e 7 listam sites e repositórios, respectivamente, onde pelo menos dois conjuntos de dados desapareceram. ScienceDirect foi o site com mais dados ausentes; metade dos 14 links do ScienceDirect estavam faltando. Muitos outros sites de periódicos aparecem na Tabela 6, sugerindo que informações suplementares hospedadas por periódicos não são um método estável de compartilhamento de dados. Como as informações suplementares são tão prevalentes (ver discussão anterior sobre Links PDF) e foram apenas incidentalmente testadas nesta análise, a disponibilidade contínua de informações suplementares hospedadas por periódicos merece investigação adicional. O GitHub também aparece no topo da Tabela 6, com 5 conjuntos de dados ausentes, sugerindo que também não é um local ideal para armazenar dados de forma sustentável.

Discussão

Esta pesquisa tem várias limitações. Primeiro, esta análise não é uma medida completa de "decomposição de dados", pois não verificou a deriva de conteúdo (ou seja, se a página da web coletada realmente representa os dados)