



ANÁLISE DOS CONJUNTOS DE DADOS DEPOSITADOS NO REPOSITÓRIO DE DADOS DE PESQUISA DA EMBRAPA (REDAPE)

Nicolas Gustavo **Ricci**¹; Debora Pignatari **Drucker**²; Carla Geovana do Nascimento **Macario**³,
Victor Paulo Marques **Simão**⁴, Juliana Meireles **Fortaleza**⁵, Patricia Rocha Bello **Bertin**⁶,
Alessandra Rodrigues da **Silva**⁷

Nº 24601

RESUMO – O Redape, Repositório de Dados de Pesquisa da Embrapa, é um repositório fundamentado nos princípios FAIR (do inglês encontráveis, acessíveis, interoperáveis e reusáveis), que facilita a preservação e acesso aos dados de pesquisa gerados pela instituição. Este estudo teve como objetivo analisar os conjuntos de dados depositados no repositório nos últimos dois anos, considerando a quantidade de conjuntos de dados depositados por categoria temática, a quantidade de conjuntos de dados publicados por ano; a quantidade de conjuntos dados publicados, em revisão ou em esboço; a quantidade de arquivos depositados; a quantidade de arquivos restritos e, por fim, as palavras-chave utilizadas para representar os assuntos, distinguidas entre termos livres e termos provenientes de vocabulários controlados. De 14 de abril de 2022 a 3 de junho de 2024, foram depositados 612 conjuntos de dados no Redape, dos quais 239 já foram publicados, além de 6.686 arquivos de dados depositados, dos quais 27% estão com acesso restrito. Foram recuperadas 880 palavras-chave, sendo 654 delas distintas. Considerando os conjuntos de dados publicados, são 373 palavras-chave, entre repetidas e distintas, das quais 41 são identificadas como termos livres. Já nos conjuntos enviados para revisão, há 348 termos livres do total de 507 palavras-chave, evidenciando a relevância do papel do curador, que faz um trabalho minucioso de adicionar termos controlados, facilitando a encontrabilidade e o reuso dos conjuntos de dados. Ao adotar essa abordagem abrangente de análise, esperamos contribuir para uma compreensão do papel do Redape na gestão e na disseminação de dados de pesquisa agrícola.

Palavras-chaves: Dados de pesquisa, Repositório de dados, Palavras-chave, Princípios FAIR, Vocabulário controlado.

1 Autor, Bolsista CNPq (PIBIT): Graduação em Engenharia Agrícola, FEAGRI / Unicamp, Campinas-SP; nicolasricci@gmail.com

2 Analista da Embrapa Agricultura Digital, Campinas, SP, debora.drucker@embrapa.br.

3 Orientadora: Pesquisadora da Embrapa Agricultura Digital, Campinas, SP; carla.macario@embrapa.br.

4 Analista da Embrapa Meio Ambiente, Jaguariúna, SP, victor.simao@embrapa.br

5 Pesquisadora da Embrapa, Brasília, DF, juliana.fortaleza@embrapa.br

6 Pesquisadora, Ouvidoria da Embrapa, Brasília - DF, patricia.bertin@embrapa.br.

7 Analista da Embrapa, Brasília, DF, alessandra.silva@embrapa.br.



ABSTRACT – *Redape, Embrapa’s Research Data Repository, is a repository based on the FAIR principles (findable, accessible, interoperable, and reusable), facilitating the preservation and access to research data generated by the institution. This study aimed to analyze the datasets deposited in the repository in the last two years, considering the number of datasets deposited by thematic category, the number of datasets published per year; the number of datasets published, under review, or in draft; the number of files deposited; the number of restricted files and, finally, the keywords used to represent the subjects, distinguished between free terms and terms from controlled vocabularies. From April 14, 2022, to June 3, 2024, 612 datasets were deposited in Redape, of which 239 have already been published, in addition to 6,686 data files deposited, of which 27% have restricted access. A total of 880 keywords were retrieved, 654 of which were distinct. Considering the published datasets, there are 373 keywords, both repeated and distinct, of which 41 are identified as free terms. In the datasets sent for review, there are 348 free terms out of a total of 507 keywords, highlighting the relevance of the role of the curator, who does a meticulous job of adding controlled terms, facilitating the findability and reuse of the datasets. By adopting this comprehensive analysis approach, we hope to contribute to understanding Redape’s role in managing and disseminating agricultural research data.*

Keywords: Search data, FAIR data, Data repository, Keywords, Controlled vocabularies.

1. INTRODUÇÃO

Os dados de pesquisa têm assumido papel crucial no cenário científico global. Atualmente, a necessidade de compartilhá-los é uma questão fundamental, que reflete o crescente reconhecimento de sua importância na comunidade científica internacional (Cavalcanti e Sales, 2017). Nesse contexto, o Redape, Repositório de Dados de Pesquisa da Embrapa¹ (Figura 1), emerge como uma ferramenta que visa preservar e facilitar a busca por dados de pesquisa gerados pela Empresa Brasileira de Pesquisa Agropecuária - Embrapa, uma das principais instituições de pesquisa do Brasil. Ele permite a organização, a descrição, o gerenciamento e a publicação de dados de acordo com os Princípios FAIR, que norteiam a gestão de dados científicos em todo o mundo, a partir de critérios relacionados à acessibilidade, interoperabilidade, reprodutibilidade e reuso dos dados (Wilkinson et al., 2016).

O Redape está estruturado em nove categorias temáticas que representam as ações de Pesquisa, Desenvolvimento e Inovação desenvolvidas pela Embrapa: ‘Produção Animal’,

¹ In: <https://www.embrapa.br/redape>, consultado em 13/06/2024



'Biomassa', 'Bioinsumos e Energia Renovável', 'Biotecnologia, Nanotecnologia, Agricultura de Precisão e Digital', 'Produção Vegetal', 'Recursos Genéticos', 'Agroecossistemas, Recursos Naturais e Meio Ambiente', 'Alimentos e Nutrição Humana', 'Economia, Desenvolvimento e Sociologia Rural' e 'Inovação Organizacional'. Essas categorias não são meramente divisões técnicas, mas entidades que representam o vasto conhecimento gerado pela Empresa — cada categoria agrupa diferentes portfólios de pesquisa, conforme detalhado no manual de uso do Redape².

The screenshot displays the Redape website interface. At the top, there is a navigation bar with links for 'BRASIL', 'CORONAVÍRUS (COVID-19)', 'Simplifique!', 'Participe', 'Acesso à Informação', 'Legislação', and 'Canais'. Below this is a search bar and a 'Pesquisa' button. The main header features the Redape logo and the text 'Repositório de Dados de Pesquisa da Embrapa'. A banner below the header shows various scientific icons. The main content area displays search results for 'Anotações de responsabilidade técnica (ART) de serviços relacionados à criação de suínos e aves no estado de Santa Catarina'. The results list includes the title, author (Conselho Regional de Engenharia e Agronomia (SC)), publication year (2023), and a brief description of the document. The interface also shows filters for 'Comunidades Dataverses', 'Conjuntos de Dados', and 'Arquivos', along with a 'Pesquisa avançada' button.

Figura 1. Interface do Redape.

Um aspecto crucial para a adequada representação de dados em repositórios é o registro de palavras-chave nos metadados, as quais são destacadas pelos subprincípios FAIR I2 ("Metadados e dados usam vocabulários que seguem os princípios FAIR") e R1 ("R1 Metadados e dados são ricamente descritos com uma pluralidade de atributos precisos e relevantes").

² In: <https://www.redape.dados.embrapa.br/guia-do-usuario.xhtml>, consultado em 13/06/2024.

Palavras-chave são termos que representam o conteúdo dos itens catalogados. Elas facilitam a recuperação da informação e aumentam a precisão na busca por estudos específicos. Embora geralmente sejam derivadas de textos em linguagem natural, é recomendável escolhê-las a partir de um vocabulário controlado³ adequado à área de pesquisa em questão. Isso garante que as palavras-chave escolhidas sejam termos padronizados, criteriosamente definidos e compilados por especialistas e reconhecidos na comunidade científica, contribuindo para uma melhor organização e identificação dos documentos, evitando a ambiguidade e aumentando a precisão nas buscas (Brandau et al., 2005).

Vocabulários controlados minimizam os problemas de comunicação entre indexadores e usuários, além de promoverem a alta recuperação e relevância dos dados (Lopes, 2002). Diversos estudos sobre o uso combinado de linguagem controlada e natural na recuperação de informações têm mostrado que essa abordagem conjunta melhora o desempenho dos resultados (Lopes, 2002). O objetivo deste trabalho é analisar os conjuntos de dados depositados no Redape, nos últimos dois anos, considerando a quantidade de conjuntos de dados depositados por categoria temática, a quantidade de conjuntos de dados publicados por ano; a quantidade de conjuntos dados publicados, em revisão ou em esboço; a quantidade de arquivos depositados; a quantidade de arquivos restritos e, por fim, as palavras-chave utilizadas para representar os assuntos, distinguidas entre termos livres e termos provenientes de vocabulários controlados. Ao adotar essa abordagem abrangente, espera-se contribuir para a compreensão do papel do Redape na gestão e na disseminação de dados de pesquisa agrícola.

2. MATERIAL E MÉTODOS

Foram analisados os conjuntos de dados registrados no Redape a partir de 14 de abril de 2022, quando o repositório foi lançado, até 3 de junho de 2024. O processo de gestão dos conjuntos de dados no Redape, descrito no Guia do Usuário, inicia com o depósito dos dados pelos empregados da Embrapa, que devem seguir as instruções para o preenchimento dos metadados obrigatórios e opcionais para que os arquivos de dados sejam adequadamente interpretáveis. Após registro dos os metadados iniciais, os depositantes salvam o conjunto de dados, que ficará armazenado no modo 'esboço', e adicionam um ou mais arquivos de dados, cujo acesso pode ser público ou restrito, de acordo com a Norma de Acesso e Tratamento da Informação da Embrapa, a qual estabelece orientações básicas para o tratamento das informações públicas e restritas na Empresa, visando assegurar níveis adequados de acesso e proteção, em alinhamento às disposições constitucionais e legais pertinentes. Uma vez completado o preenchimento de

³ A linguagem controlada, ou vocabulário controlado, consiste em uma lista de termos autorizados e pré-selecionados para organizar o conhecimento, visando a precisão na recuperação dos documentos.



metadados e a carga de arquivos de dados, o depositante submete o conjunto de dados para revisão pelos curadores, que analisam todos os elementos do conjunto de dados (metadados, arquivos, termos de uso, permissões, miniaturas) e sugerem alterações ao depositante ou as implementam diretamente, quando pertinente.. Como exemplo, caso os depositantes não tenham utilizado termos provenientes de vocabulários controlados, os curadores verificam se há termos controlados que representam conceitos equivalentes aos indicados pelos depositantes e os inserem. Os termos livres são mantidos caso não tenham sido encontrados equivalentes em vocabulários controlados.

Os conjuntos de dados salvos no modo esboço ou submetidos à revisão têm os metadados visíveis para o público interno da Embrapa, bem como os arquivos de dados caso não tenham restrição de acesso. Quando não há mais ajustes a serem feitos, os conjuntos de dados são publicados pelos curadores e os metadados ficam sempre visíveis, tanto para o público interno quanto externo, sendo que os arquivos de dados podem estar restritos ou públicos.

Os conjuntos de dados depositados no Redape foram analisados em relação à categoria temática; ano de publicação; situação (publicados, em revisão ou em esboço); quantidade de arquivos depositados; quantidade de arquivos restritos e, por fim, palavras-chave utilizadas para representar o conhecimento, distinguindo entre termos livres e termos provenientes de vocabulários controlados.

3. RESULTADOS E DISCUSSÃO

De 14 de abril de 2022 a 3 de junho de 2024, foram depositados 612 conjuntos de dados no Redape. Os conjuntos de dados estão distribuídos nas nove categorias temáticas, sumarizados a seguir e descritos na Tabela 1:



Tabela 1. Quantidade de conjuntos de dados depositados no repositório Redape, por situação (esboço, em revisão ou publicado) e tema.

Temas dos conjuntos de dados	Esboço	Em revisão	Publicado
Produção Animal	62	17	4
Biomassa, Bioinsumos e Energia Renovável	26	11	14
Biotecnologia, Nanotecnologia, Agricultura de Precisão e Digital	82	39	14
Produção Vegetal	66	31	11
Recursos Genéticos	20	10	1
Agroecossistemas, Recursos Naturais e Meio Ambiente	158	91	188
Alimentos e Nutrição Humana	5	3	0
Economia, Desenvolvimento e Sociologia Rural	1	0	4
Inovação Organizacional	1	0	3
Total	421	202	239

Esses resultados revelaram que algumas categorias temáticas foram mais selecionadas do que outras. A categoria Agroecossistemas, Recursos Naturais e Meio Ambiente tiveram 188 conjuntos de dados publicados, uma quantidade maior do que todas as outras categorias somadas. Isso se deve ao fato de que houve um esforço concentrado para transferir os dados de um projeto específico, denominado Paisagens Sustentáveis, de um repositório obsoleto para o Redape, conforme descrito por Ricci et al., (2023). Se desconsiderarmos os conjuntos de dados do projeto Paisagens Sustentáveis, apenas outros 15 conjuntos de dados foram publicados nessa categoria. Ainda assim, essa foi a categoria mais selecionada no período pesquisado, seguida por Biotecnologia, Nanotecnologia, Agricultura de Precisão e Digital; Produção Vegetal e Produção

Animal, respectivamente. As categorias menos selecionadas foram Inovação Organizacional; Economia, Desenvolvimento e Sociologia Rural e Alimentos e Nutrição Humana.

No mesmo período, foram depositados 6.686 arquivos de dados, na seguinte distribuição por status de publicação: 4.054 arquivos no modo esboço, sendo que 2.787 deles já submetidos a revisão, e 2.632 arquivos publicados. Do total de 6.686 arquivos de dados, 1.805 são de acesso restrito, o que representa cerca de 27%. As justificativas mais comuns para a restrição dos dados estão relacionadas com a proteção de dados pessoais, a propriedade intelectual ou embargo de pesquisas ainda em andamento.

Na Figura 2 apresentamos a quantidade de conjuntos de dados publicados por ano.

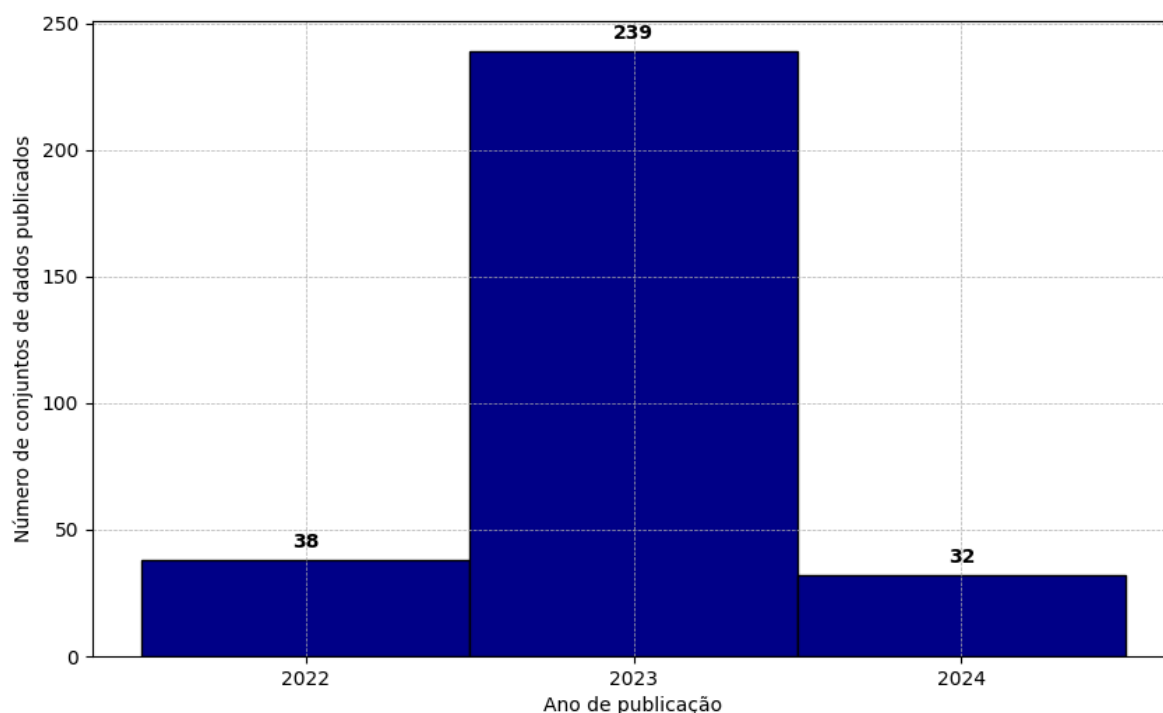


Figura 2. Quantidade de conjuntos de dados publicados no Redape de abril de 2022 a junho de 2024.

Durante os anos de 2022, 2023 e 2024, foram publicados respectivamente 38, 239 e 32 conjuntos de dados no repositório, evidenciando o crescente aumento de publicações desde o surgimento do repositório, uma vez que no ano de surgimento há 38 publicações, no ano seguinte 239 e neste ano, as publicações já quase superam o primeiro ano em apenas 6 meses. Considerando o viés dos dados do Projeto Paisagens Sustentáveis, publicados majoritariamente em 2023, e a quantidade de dados depositados em 2024 e ainda em revisão, pode-se afirmar que a quantidade de dados depositados no repositório está crescendo.



Já em relação aos conjuntos de dados depositados, ou seja, englobando tanto os esboços como os em revisão e os publicados, foram depositados 253 conjuntos de dados em 2022, ano de lançamento do Redape, dos quais 240 conjuntos são do projeto Paisagens Sustentáveis; 212 conjuntos de dados depositados em 2023, sendo 15 do projeto Paisagens Sustentáveis e, no ano atual, 2024, 147 conjuntos de dados. Considerando que até junho de 2024 foram depositados 147 conjuntos de dados, o número baixo de depósitos neste ano não significa menor procura, já que em 2023 foram depositados 212 conjuntos de dados em doze meses, e sim indica que muitos conjuntos de dados ainda estão no modo esboço ou em revisão.

Em relação às palavras-chave, foram identificadas aquelas mais recorrentes nos metadados dos conjuntos de dados hospedados no Redape. Constituem a maior parte das palavras-chave presentes no Redape os termos designados no projeto Paisagens Sustentáveis Brasil, sendo elas: "forest cover" (184 ocorrências), "forest degradation" (184 ocorrências), "LIDAR" (145 ocorrências), todas provenientes do vocabulário controlado Agrovoc. Conforme mencionado anteriormente, a transferência dos conjuntos de dados descrita por Ricci et al. (2023) explica esse resultado. Como a catalogação dos dados se deu em um contexto específico de um projeto, realizamos a análise dos termos das palavras-chave também desconsiderando os conjuntos de dados do Paisagens Sustentáveis.

Considerando os demais conjuntos de dados do repositório, detectamos o total de 880 palavras-chave, sendo 654 distintas. Ao se observar os dados publicados, obtém-se 373 palavras-chave, entre repetidas e distintas, das quais 41 são identificadas como termos livres. Os termos mais comuns, considerando tanto os conjuntos de dados publicados quanto os submetidos à revisão, foram "*Anacardium occidentale*" (nome científico do cajueiro, 9 ocorrências), sendo classificados em 8 das ocorrências como vocabulário controlado e 1 como termo livre, "Cana de Açúcar" (8 ocorrências), 5 classificados como vocabulário controlado e 3 como termo livre, "Caju" (7 ocorrências), todas ocorrências classificadas como vocabulário controlado, "Bovino de corte" (7 ocorrências), todos classificadas como termo livre, "Melhoramento de Plantas" (7 ocorrências), todos classificados como vocabulário controlado, "Agrometeorologia" (6 ocorrências), quatro classificados como termo livre e dois como vocabulário controlado, "Fenótipo" (6 ocorrências), todas classificadas como termo livre, e "Plant breeding" (6 ocorrências), classificadas como vocabulário controlado. Percebe-se que as palavras-chave de maior ocorrência estão relacionadas às categorias temáticas "Produção animal" e "Produção vegetal", seguidas por "Recursos Genéticos" com as palavras-chave melhoramento de plantas, fenótipo, plant breeding e, finalmente, "Agroecossistemas, Recursos Naturais e Meio Ambiente" com a palavra-chave agrometeorologia.

Conforme já mencionado, os dados são depositados no Redape pelos próprios autores e são revisados por curadores. Considerando os conjuntos de dados em revisão, que somam 507 palavras-chave, dentre elas, 348 não contém palavras-chave de vocabulários controlados. Durante



a revisão, os curadores adicionam palavras-chave de vocabulários controlados nos metadados e mantém apenas termos livres que não foram encontrados nos principais vocabulários controlados da área de ciência agropecuária, como o Agrovoc⁴ e o NAL⁵, assegurando que os dados publicados estejam de acordo com os padrões estabelecidos. Esses resultados indicam que as palavras-chave podem ser melhor exploradas no Redape, tanto na capacitação e orientação aos depositantes a utilizar os vocabulários controlados quanto para facilitar a recuperação de dados, bem como para permitir que se realizem estudos mais detalhados sobre os temas dos trabalhos cujos dados estão depositados nesse repositório.

4. CONCLUSÃO

No período analisado, que abrange pouco mais de dois anos, centenas de conjuntos de dados foram depositados no Redape. Considerando que mais da metade ainda está em modo “esboço”, observou-se a necessidade de incentivar que os conjuntos de dados não submetidos à revisão sejam finalizados e submetidos, o que pode ser fomentado por meio de estratégias de comunicação e sensibilização.

Além disso, os resultados indicam que é necessária maior disseminação e conscientização da importância do uso de vocabulários controlados, considerando que os depositantes utilizaram pouco essas ferramentas. Isso evidencia também que a curadoria desempenha um papel fundamental no processo, em especial por contribuir para que os conjuntos de dados atendam aos padrões estabelecidos, o que facilita sua encontrabilidade e reúso. A compreensão desses aspectos é fundamental para melhorar a gestão e a disseminação dos dados de pesquisa agropecuários, promovendo o uso eficiente desses ativos. Ao adotar essa abordagem abrangente de análise, esperamos contribuir para uma compreensão do papel do Redape na gestão e na disseminação de dados de pesquisa agrícola.

5. AGRADECIMENTO

Agradeço ao CNPQ pela concessão da bolsa PIBIT e à Embrapa Agricultura Digital.

⁴ In: <https://agrovoc.fao.org/browse/agrovoc/en/>, consultado em: 13/06/2024.

⁵ In: <https://agclass.nal.usda.gov/form/thesaurus-search>, consultado em: 13/06/2024.



6. REFERÊNCIAS

BIBLIOTECA esclarece a diferença entre palavras-chave e descritores. 202. Disponível em: <<https://www.ifsudestemg.edu.br/noticias/sidr/2021/07/biblioteca-esclarece-a-diferenca-entre-palavras-chave-e-descritores>>. Acesso em: 20 jun. 2024.

BRANDAU, RICARDO; MONTEIRO, ROSANGELA; BRAILE, Domingo M. Importância do uso correto dos descritores nos artigos científicos. Revista Brasileira de Cirurgia Cardiovascular, São José do Rio Preto, v. 20, n. 1, p. 7-9, jan./mar. 2005. DOI: <https://doi.org/10.1590/S0102-76382005000100004>. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-76382005000100004&lng=en&nrm=iso. Acesso em: 25 jun. 2021.).

CAVALCANTI, M. T.; SALES, L. F. Gestão de dados de pesquisa: um panorama da atuação da União Europeia. Biblos, Rio Grande, v. 31, n. 1, p. 73-98, ago. 2017. DOI: <https://doi.org/10.14295/biblos.v31i1.5789>. Disponível em: <https://periodicos.furg.br/biblos/article/view/5789>. Acesso em: 10 jun. 2024.

LOPES, I. L.. (2002). Uso das linguagens controlada e natural em bases de dados: revisão da literatura. Ciência Da Informação, 31(1), 41–52. <https://doi.org/10.1590/S0100-19652002000100005>.

REDAPE. Disponível em: <https://www.redape.dados.embrapa.br/>. Acesso em: 13 jun. 2024.

RICCI, N. G.; DRUCKER, D. P.; VICTORIA, D. de C.; MACARIO, C. G. do N. Princípios FAIR aplicados à gestão de dados de inventário florestal e LiDAR: o caso do projeto paisagens sustentáveis.

WILKINSON MD, DUMONTIER M, AALBERSBERG IJ, APPLETON G, AXTON M, BAAK A, BLOMBERG N, BOITEN JW, DA SILVA SANTOS LB, BOURNE PE, BOUWMAN J, BROOKES AJ, CLARK T, CROSAS M, DILLO I, DUMON O, EDMUNDS S, EVELO CT, FINKERS R, GONZALEZ-BELTRAN A, GRAY AJ, GROTH P, GOBLE C, GRETHE JS, HERINGA J, 'T HOEN PA, HOOFT R, KUHN T, KOK R, KOK J, LUSHER SJ, MARTONE ME, MONS A, PACKER AL, PERSSON B, ROCCA-SERRA P, ROOS M, VAN SCHAIK R, SANSONE SA, SCHULTES E, SENGSTAG T, SLATER T, STRAWN G, SWERTZ MA, THOMPSON M, VAN DER LEI J, VAN MULLIGEN E, VELTEROP J, WAAGMEESTER A, WITTENBURG P, WOLSTENCROFT K, ZHAO J, MONS B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. doi: 10.1038/sdata.2016.18. Erratum in: Sci Data. 2019 Mar 19;6(1):6. PMID: 26978244; PMCID: PMC4792175.