# Repositórios de Dados Científicos na Infraestrutura de Pesquisa: adoção dos princípios FAIR

**Elizabete Cristina de Souza de Aguiar Monteiro**
Mestre, Universidade Estadual Paulista (UNESP), Marília, SP, Brasil.
Doutoranda do Programa de Pós-Graduação em Ciência da Informação (UNESP), Marília, SP, Brasil.
http://lattes.cnpq.br/3258820169472861
ecsamonteiro@gmail.com

**Ricardo César Gonçalves Sant'Ana**
Livre-docente, Universidade Estadual Paulista (UNESP), Marília, SP, Brasil.
Docente do Programa de Pós-Graduação em Ciência da Informação (UNESP), Marília, SP, Brasil.
http://lattes.cnpq.br/1022660730972320
ricardo.santana@unesp.br

Data de submissão: 12/09/2019. Data de aprovação no ConfOA: 12/06/2019. Data de publicação:

**RESUMO**

As infraestruturas de pesquisa são implementadas por instituições para apoiar a descoberta, a partilha, a exploração, a análise, a curadoria, a utilização, a replicação e a gestão de dados que, alinhados aos aspectos debatidos no contexto da Ciência Aberta, têm o potencial de acelerar a descoberta e a inovação científica. Os princípios FAIR para a ciência de dados estão sendo adotados em conexão com infraestruturas de pesquisa como CLARIN, que fornecem repositórios de dados para sua própria gestão. O Objetivo do artigo é apresentar a infraestrutura de pesquisa CLARIN, a qual adotou os princípios FAIR, e discutir a relevância dos repositórios de dados nesse contexto. Foi utilizada a metodologia qualitativa como base do levantamento bibliográfico para a discussão sobre as temáticas. Para coleta sobre os repositórios de dados da infraestrutura CLARIN foi utilizado o Registry of Research Data Repositories (re3data.org). Conclui-se que os repositórios de dados propiciam condições para atenderem aos requisitos do FAIR e que a quantidade de repositórios de dados integrados à infraestrutura CLARIN demonstra o valor agregado desses repositórios.

***Palavras-chave:*** Repositório de dados. Dados científicos. Infraestrutura de pesquisa. Princípios FAIR. Acesso aberto.

# *Scientific Data Repositories in Research Infrastructure: adoption of FAIR Principles*

***ABSTRACT***

*Research infrastructures are implemented by institutions to support the discovery, sharing, exploration, analysis, curation, use, replication and management of data that, line up with the issues discussed in the Open Science context, have the potential accelerate discovery and scientific innovation. FAIR principles for data science are being adopted in connection with research infrastructures such as CLARIN that provide data repositories for data management. The purpose of this paper is to present the CLARIN research infrastructure which adopted the FAIR principles and discuss the relevance of data repositories in this context. The qualitative methodology was used as base the bibliographic survey for the discussion about the themes. To collect data on CLARIN infrastructure data, the Registry of Research Data Repositories (re3data.org) was used. It is concluded that data repositories provide conditions to accord FAIR requirements and that the number of data repositories integrated into the CLARIN infrastructure demonstrates the added value of these repositories.*

***Keywords****: Data repository. Scientific data. Research Infrastructure. FAIR Principles. Open access.*

# *Repositorios de datos científicos en infraestructura de investigación: adopción de principios FAIR*

***RESUMEN***

*Las instituciones implementan las infraestructuras de investigación para apoyar el descubrimiento, el intercambio, la exploración, el análisis, la curación, el uso, la replicación y la gestión de datos que, de acuerdo con los temas discutidos en el contexto de Open Science, tienen el potencial de acelerar el descubrimiento y la innovación científica. Los principios FAIR para la ciencia de datos se están adoptando en relación con las infraestructuras de investigación, como CLARIN, que proporcionan depósitos de datos para la gestión de datos. El propósito de este documento es presentar la infraestructura de investigación CLARIN que adoptó los principios FAIR y discutir la relevancia de los repositorios de datos en este contexto. Se utilizó la metodología cualitativa como base de la encuesta bibliográfica para la discusión sobre los temas. Para recopilar datos sobre los datos de la infraestructura CLARIN, se utilizó el Registro de repositorios de datos de investigación (re3data.org). Se concluye que los repositorios de datos proporcionan condiciones para cumplir con los requisitos FAIR y que el número de repositorios de datos integrados en la infraestructura CLARIN demuestra el valor agregado de estos repositorios.*

***Palabras clave:*** *Repositorio de datos. Datos cientificos. Infraestructura de investigación. Principios FAIR. Acceso abierto.*

## INTRODUÇÃO

As Infraestruturas no contexto da pesquisa são sistemas de grande escala, com múltiplos componentes sociais e técnicos, incluindo recursos humanos, trabalho colaborativo, aparatos tecnológicos, acesso distribuído, informação documentada, instituições e repositórios (BORGMAN *et al.*, 2015; EDWARDS *et al.*, 2007; EDWARDS *et al.*, 2013).

O Common Language Resources and Technology Infrastructure (CLARIN) é uma infraestrutura de pesquisa que tem sua gênese em um projeto colaborativo de países europeus com os objetivos de garantir acesso, integração e exploração de grandes quantidades de dados linguísticos e o uso de ferramentas tecnológicas na pesquisa em áreas de humanidades e ciências sociais (BEL *et al.*, 2008; IRUSKIETA, 2016; DE SMEDT *et al.*, 2018).

A infraestrutura CLARIN suporta o compartilhamento, o uso e a sustentabilidade de dados e ferramentas de idiomas por meio de uma federação de centros em rede que compreendem repositórios de dados de idiomas, centros de serviços e centros de conhecimento (FIŠER; LENARDIČ; ERJAVEC, 2018).

Os pilares técnicos da infraestrutura CLARIN são: identidade federada, identificadores persistentes, repositórios sustentáveis, metadados flexíveis e definições de conceitos, pesquisa de conteúdo e encadernação de serviços da web (CLARIN, 2019). "Os fatores importantes para o sucesso e sustentabilidade de uma infraestrutura de pesquisa como CLARIN são seu escopo, tamanho e estrutura. O CLARIN lida com dados de linguagem digital e sua curadoria e processamento" (DE SMEDT *et al.*, 2018, p. 2, tradução nossa).

O CLARIN torna realidade a visão que está subjacente às políticas europeias emergentes e aos paradigmas em relação à Ciência Aberta, interconectando pesquisadores por meio das fronteiras disciplinares, nacionais e internacionais, oferecendo gestão e acesso contínuo a dados e a serviços on-line e disponibilizando os dados sob os princípios FAIR — Findable, Accessible, Interoperable, Reusable (DE SMEDT *et al.*, 2018; DE JONG *et al.*, 2018). Vale destacar que nem todos os dados abertos, especialmente dados científicos, são ou devem ser completamente "abertos" ou "gratuitos", mas devem pelo menos ser FAIR (EUROPEAN COMMISSION, 2018).

O FAIR oferece um conjunto de princípios para aprimorar a utilidade dos dados e marca um refinamento importante dos conceitos necessários para dar maior valor aos dados e aumentar sua propensão para reutilização, por seres humanos e por máquinas, auxiliando na interação entre quem deseja usar os dados e quem os fornece (KALINAUSKAITĖ, 2017; EUROPEAN COMMISSION, 2018). Ademais, exige grandes mudanças em termos de cultura e prática de pesquisa, além de implementação de um ecossistema com serviços e componentes de dados como Políticas, Plano de Gestão e Dados, Identificadores, Padrões e Repositórios, de modo que esse último é essencial no ecossistema de dados FAIR, pois é necessário para executar a função de oferecer dados e metadados acessíveis e reutilizáveis para usuários e por máquinas (EUROPEAN COMMISSION, 2018).

Os centros CLARIN, em geral, fornecem repositórios de dados com infomações de pesquisa que podem ser os resultados de projetos, de grupos de pesquisa ou de estudiosos individuais (DE SMEDT *et al.*, 2018). Com a gestão dos dados nos repositórios é possível acessar a disponibilização de conjuntos de dados, de ferramentas e de recursos neles armazenados, além de publicação de dados em FAIR.

Os repositórios de dados

> [...] tem sua gênese com a necessidade de gestão dos dados científicos, estão vinculados às universidades e instituições de pesquisa e contribuem para assegurar que os dados sejam publicados e disponibilizados para a comunidade científica com o menor número possível de restrições (MONTEIRO; SANT'ANA, 2017).

Este artigo consiste em apresentar a infraestrutura de pesquisa CLARIN, que adotou os princípios FAIR e demonstra a relevância e o papel estratégico dos repositórios de dados nesse contexto.

## MATERIAL E MÉTODOS

A metodologia é qualitativa e tem como base o levantamento bibliográfico para a contextualização e discussão sobre as temáticas. Para o levantamento das informações sobre os repositórios de dados da infraestrutura CLARIN foi utilizado o Registry of Research Data Repositories (re3data.org)[1], um registro global de repositórios de dados de pesquisa que abrange repositórios de diferentes disciplinas.

Após o levantamento dos repositórios que compõem a infraestrutura CLARIN, foi utilizada a técnica checklist com o instrumento de coleta itens de checagem para verificação de quais dos princípios FAIR já foram adotados pelos repositórios. A coleta dos itens do checklist foram feitas no re3data.org, que é o site do CLARIN e dos repositórios.

Os critérios que compuseram o item de checagem foram os princípios FAIR (Findable, Accessible, Interoperable, Reusable), em que são estabelecidos subprincípios em relação aos dados e aos metadados (EUROPEAN COMMISSION, 2018). Os resultados foram apresentados em um quadro. O checklist deste artigo foi organizado em conformidade com a lista de itens que compõem os princípios FAIR e seus subprincípios.

## RESULTADOS E DISCUSSÕES

O CLARIN é uma infraestrutura distribuída que possui 20 membros. É composto pela Áustria, Bulgária, República Checa, Dinamarca, União de Língua Holandesa, Estônia, Finlândia, Alemanha, Grécia, Hungria, Itália, Letônia, Lituânia, Holanda, Noruega, Polônia, Portugal, Eslovênia, Suécia, por dois observadores (França e Reino Unido) e um país (Estados Unidos da América) (CLARIN, 2019). A busca no re3data.org recuperou 58 repositórios.

Um dos serviços fundamentais da infraestrutura CLARIN é garantir que os recursos linguísticos possam ser arquivados e disponibilizados à comunidade de maneira confiável e sustentável. Nesse sentido, muitos dos centros CLARIN oferecem repositórios de dados para o armazenamento dos conjuntos de dados (CLARIN, 2019).

Os repositórios são integrados à infraestrutura do CLARIN e proporcionam ampla exposição dos conjuntos de dados depositados e de softwares para análise, processamento e visualização dos dados, colaborando para a visibilidade e para o trabalho colaborativo dos pesquisadores.

Observou-se que cada repositório está em uma fase de desenvolvimento e localizado em instituições de países diferentes. Há uma diversidade de conjuntos de dados e ferramentas que variam, sendo textos escritos e falados, registros de áudios e vídeos, recursos e ferramentas lexicais, textos estruturados com anotações, ferramentas para trabalhar com eles que vão desde analisadores linguísticos até ambientes de programação, bancos de dados lexicais, corporações de texto, corporações de fala, ferramentas de tecnologia de fala, entre outros.

O papel estratégico dos repositórios de dados favorece a disponibilização dos dados e apoia a infraestrutura CLARIN nos princípios FAIR, corroborando suas potencialidades (QUADRO 1).

[1] https://www.re3data.org/

Quadro 1 – Cheklist Princípios FAIR

| Princípio | Recomendações | Repositórios CLARIN |
|---|---|---|
| Findable | F1- (meta)dados são atribuídos a um identificador globalmente exclusivo e persistente | Uso de identificadores persistentes para dados |
| | F2 - os (meta)dados são descritos com metadados ricos (definidos por R1 no princípio Reusable) | Requer o uso do Component MetaData Infrastructure (CMDI); uso do Dublin Core |
| | F3 - (meta)dados de forma clara e explícita incluem o identificador dos dados que descreve | Uso de handler e identificadores persistentes |
| | F4 - (meta)dados são registrados ou indexados em um recurso pesquisável | Catálogo de registros de metadados Virtual Language Observatory (VLO); testes de usabilidade e melhoria; curadoria dos metadados |
| Accessible | A1 - (meta)dados são recuperáveis pelo seu identificador usando um protocolo de comunicação padronizado | Protocolos: HTTP; SAML; OAI-PMH |
| | A1.1 - o protocolo é aberto, gratuito e universalmente implementável | Protocolos: HTTP; SAML; OAI-PMH |
| | A1.2 - o protocolo permite um procedimento de autenticação e autorização, quando necessário | Protocolos: HTTP e SAML |
| | A2 - os (meta)dados estão acessíveis, mesmo quando os dados não estão mais disponíveis | Sim |
| Interoperable | I1 - Os (meta)dados usam uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento | Estrutura CMDI como linguagem de metadados |
| | I2 - (meta)dados usam vocabulários que seguem os princípios do FAIR | Links para o vocabulário OpenSKOS |
| | I3 - Os (meta)dados incluem referências qualificadas a outros (meta) dados | Sim |
| Reusable | R1 meta(data) são ricamente descritos com uma pluralidade de atributos precisos e relevantes | Uso do CMDI e Dublin Core |
| | R1.1 - (meta)dados são liberados com uma licença de uso de dados clara e acessível | Uso das licenças Creative Commons; Open Data Common; Apache License 2.0 |
| | R1.2 - (meta)dados estão associados à proveniência detalhada | Sim |
| | R1.3 - (meta)dados atendem aos padrões da comunidade relevantes ao domínio | Sim |

Fonte: Elaborado pelos autores baseados em European Commission (2019).

Ainda que os princípios FAIR se aplicam principalmente aos meta(dados), sua implementação requer vários serviços de dados e componentes a serem implementados que são implementáveis com os repositórios de dados (EUROPEAN COMMISSION, 2018). As potencialidades do apoio técnico dos repositórios na infraestrutura CLARIN se destacam nos seguintes casos:

- arquivamento a longo prazo: armazenamento pode ser concedido por um longo período (até 50 anos em alguns casos);
- curadoria dos conjuntos de dados;
- disponibilização dos conjuntos de meta(dados) e recursos de linguagens aos pesquisadores;
- exposição dos conjuntos de dados colaboram para a visibilidade e para o trabalho colaborativo dos pesquisadores;
- informações sobre como e sob quais condições os conjuntos de dados e os recursos existentes podem ser reutilizados;
- nos conjuntos de dados é indicada uma licença e informações de proveniência e aderência aos padrões da comunidade;
- os recursos podem ser citados facilmente com um identificador persistente;
- os metadados dos recursos e dos dados são coletados e indexados pelo Virtual Language Observatory (VLO);
- recursos protegidos por senha podem ser disponibilizados por meio de um login institucional;
- os recursos são integrados à infraestrutura CLARIN e podem ser analisados e enriquecidos mais facilmente com várias ferramentas linguísticas (KALINAUSKAITĖ, 2017; DE JONG et al., 2018).

Os repositórios são componentes essenciais no ecossistema FAIR (EUROPEAN COMMISSION, 2018). Os repositórios gerenciam o acesso a dados e metadados e oferecem serviços para acesso e reutilização.

A gestão de dados, assim como tornar os dados FAIR, em muitos casos não faz parte da prática de pesquisadores individuais ou pequenos.

Adaptar métodos científicos a maiores volumes de dados, muitas vezes com maior diversidade, apresenta novos desafios para a ciência e para o gestão de dados (BORGMAN *et al.*, 2015). A combinação de dados de múltiplas pesquisas para análise colaborativa e novas interpretações requer sistemas e serviços que são estruturados por infraestruturas colaborativas, corroborando a necessidade da implementação de repositórios de dados para apoiar o ciclo de vida dos dados.

A implementação de repositórios para apoiar infraestruturas deve ser flexível o suficiente para suportar diferentes metodologias de colaboração e apoiar a adição de metodologias de colaboração após a implantação do sistema — por exemplo, quando o repositório de dados já está preenchido por dados de projetos (INDRUSIAK; GLESNER; REIS, 2002). Os repositórios são essenciais para infraestruturas de pesquisa com ecossistema de dados FAIR, pois são necessários para desempenhar a função de oferecer dados e metadados acessíveis e reutilizáveis aos usuários (EUROPEAN COMMISSION, 2018).

## CONSIDERAÇÕES FINAIS

A quantidade de repositórios de dados integrados à infraestrutura CLARIN assevera o potencial dos repositórios de dados na gestão e publicação de conjuntos de dados. Os repositórios da CLARIN disponibilizam dados e metadados estruturados, registram a proveniência dos dados e propiciam a interoperabilidade entre os sistema, sem esquecer-se do direito autoral no acesso e reuso dos dados. Os princípios FAIR definem as características que os recursos, as ferramentas, os vocabulários e a infraestrutura de dados devem exibir para contribuir com a descoberta e reutilização de dados e metadados. A infraestrutura CLARIN viabiliza o trabalho colaborativo e a gestão de dados compartilhados, além da reprodutibilidade de conteúdo já mapeado e analisado que, seguindo os princípios FAIR, participa do paradigma da Ciência Aberta.

## REFERÊNCIAS


BEL, N.; GONZÁLEZ-BLANCO GARCÍA, E.; IRUSKIETA, M. CLARIN Centro -K-español. *Procesamiento del Lenguaje Natural*, Alicante, n. 57, p. 151-154, 2016. Disponible en: http://www.redalyc.org/articulo.oa?id=515754424019. Acesso em: 17 mar. 2019.

BEL, N. *et al*. El proyecto CLARIN: una infraestructura de investigación científica para las humanidades y las ciencias sociales. *Digithum*: les humanitats en l'era digital, Catalunya, n. 10, p. 1-8 maio 2008. Disponível em: https://dialnet.unirioja.es/servlet/articulo?codigo=4805587. Acesso em: 3 jan. 2017.

BORGMAN, C. L. *et al*. Knowledge infrastructures in science: data, diversity, and digital libraries. *International Journal on Digital Libraries,* New York, v. 16, n. 3, p. 207–227, sep. 2015. Disponível em: http://link.springer.com/article/10.1007%2Fs00799-015-0157-z#page-1. Acesso em: 2 nov. 2015.

CLARIN. *Participating Consortia*. Netherlands. Disponível em https://www.clarin.eu/content/participating-consortia. Acesso em: 2 nov. 2019.

DE SMEDT, K. *et al*. Towards an open science infrastructure for the digital humanities: the case of CLARIN. *In*: CONFERENCE ON DIGITAL HUMANITIES IN THE NORDIC COUNTRIES, 3., 2018, Helsinki. *Proceedings* [...]. Aachen: CEUR Workshop Proceedings, 2018. p. 139-151. Disponível em: http://ceur-ws.org/Vol-2084/paper11.pdf. Acesso em: 07 mar. 2019.

DE JONG, F. M. G. *et al*. CLARIN: Towards FAIR and responsible data science using language resources. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2018), 11., 2018, Miyazaki. *Proceedings*[…]. Utrecht: Utrecht University Repository, 2018. p. 3259 - 3264. Disponível em: https://dspace.library.uu.nl/handle/1874/364776. Acesso em: 07 mar. 2019.

DOWNEY, L.; BANERJEE, S. Building an information architecture checklist: encouraging and enabling IA from infrastructure to the user interface architecture. *Journal of Information Architecture*, [*S.l.*], v. 2, n. 2, p. 25-42, 2010.

EDWARDS, P. N. *et al. Understanding infrastructure*: dynamics, tensions and design. Ann Arbor: Deep Blue, 2007. Disponível em: https://deepblue.lib.umich.edu/handle/2027.42/49353. Acesso em: 2 nov. 2015.

EDWARDS, P. N. *et al. Knowledge infrastructures*: intellectual frameworks and research challenges. Ann Arbor: Deep Blue, 2013. Disponível em: http://pne.people.si.umich.edu/PDF/Edwards_etal_2013_Knowledge_Infrastructures.pdf. Acesso em: 2 nov. 2015.

EUROPEAN COMMISSION. *Turning FAIR into reality*. Luxembourg: Publications Office of the European Union, 2018. Disponível em: https://publications.europa.eu/en/publication-detail/-/publication/d. /language-en/format-PDF/source-80611283. Acesso em: 20 dez. 2018.

FIŠER,D.; LENARDIČ, J.; ERJAVEC, T. CLARIN's Key Resource Families. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC-2018), 11., 2018, Miyazaki. *Proceedings* [...]. França: European Languages Resources Association, 2018. Disponível em: https://www.aclweb.org/anthology/L18-1210. Acesso em: 7 maio 2019.

INDRUSIAK, L. S.; GLESNER, M.; REIS, R. Comparative analysis and application of data repository infrastructure for collaboration-enabled distributed design environments. *In*: CONFERENCE ON DESIGN, AUTOMATION AND TEST IN EUROPE, 2., 2002. *Proceedings* […]. Washington: IEEE Computer Society, 2002. Disponível em: https://dl.acm.org/citation.cfm?id=874419. Acesso em: 29 jun. 2019.

KALINAUSKAITĖ, D. *To be findable, accessible, interoperable and reusable*: language data and technology infrastructure for supporting the FAIR data approach. [*S.l.*], 2017. Disponível em: https://pdfs.semanticscholar.org/fe23/3a6acf062719d9834d4ba71d8aa0acee82b4.pdf?_ga=2.154795704.1673960049.1568136138-885200410.1566584314. Acesso em: 29 jun. 2019.

MONTEIRO, E. C. S. A.; SANT'ANA, R. C. G. Plano de gerenciamento de dados em repositórios de dados de universidades. *Encontros Bibli*, Florianópolis, v. 23, n. 53, p. 160-173, set./dez. 2018. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2018v23n53p160/37296. Acesso em: 10 jan. 2019.