Big Data Architectures for FAIR-compliant Repositories: A Systematic Review

João P. C. Castro[1,2], **Cristina D. Aguiar**[1]

[1]Department of Computer Science – University of São Paulo – Brazil

[2]Information Technology Board – Federal University of Minas Gerais – Brazil

jpcarvalhocastro@ufmg.br, cdac@icmc.usp.br

Abstract. *The FAIR Principles state that scientific data should be Findable, Accessible, Interoperable, and Reusable in order to adhere to the Open Science movement. However, designing a FAIR-compliant repository can be a challenge due to the complexity of managing a huge volume and variety of research data and metadata, which can also be generated at a high velocity. This complexity calls for a Software Reference Architecture (SRA) to guide data engineers during the implementation process. In this paper, we conduct a systematic review that encompasses research efforts regarding architectural solutions for implementing FAIR-compliant repositories. We analyze 323 references from Scopus, ACM, IEEEXplore, and specialists recommendations. From this analysis, we discover 7 studies that describe general purpose big data SRAs, 13 pipelines that implement the FAIR Principles to specific contexts, and 3 FAIR-compliant big data SRAs. We describe their key characteristics and discuss their limitations, highlighting tendencies and research opportunities.*

1. Introduction

The concept of Open Science has emerged in the scientific community to increase collaboration between researchers across the globe. It states that every digital asset originated from research objects should be made available and usable free of charge [Medeiros et al. 2020]. To standardize the development of data sharing repositories capable of adhering to the Open Science concept, the FAIR Principles have been proposed [Wilkinson et al. 2016]. The objective behind these principles resides in ensuring that the aforementioned digital assets are findable, accessible, interoperable, and reusable by both humans and machines. However, their implementation might be challenging depending on the volume, variety, and velocity of the scientific data and metadata to be shared, which is a complexity inherent to big data environments [Chen et al. 2014].

Considering this complexity and the fact that the FAIR Principles are defined in proximity to the user level, a data engineer would benefit considerably from adopting a Software Reference Architecture (SRA) during the implementation process. An SRA can be defined as an architectural framework that encapsulates the expertise on creating specific system architectures (or pipelines) within a particular domain [Nakagawa et al. 2011]. Consequently, a FAIR-compliant SRA would serve as a guiding blueprint for data engineers when constructing a big data sharing repository, effectively connecting the FAIR Principles with specific implementation details.

Given the importance of big data FAIR-compliant SRAs to the context of Open Science, we conduct a systematic review of the literature encompassing this domain. A

systematic review is a rigorous method that systematically searches, selects, appraises, and synthesizes existing research studies. It aims to provide an evidence-based summary of specific research questions, following predefined criteria to minimize bias while identifying research gaps [Scannavino et al. 2017]. In the literature, the work of Davoudian and Liu (2020) surveys big data SRAs available up to the year of 2020. However, the authors do not consider the FAIR principles in their comparisons. In our work, we analyze different architectural solutions for implementing big data sharing repositories capable of fulfilling the FAIR Principles. To the best of our knowledge, our work is novel since no other survey covering this specific type of analysis has been returned by the search engines during the conduction of the systematic review.

Our paper presents the following contributions:

- A systematic review of the literature that employs a reproducible methodology, laying the groundwork for future research on FAIR-compliant big data SRAs.
- A synthesis of architectural solutions for implementing repositories in line with the FAIR Principles. We also classify these solutions as general purpose big data SRAs, context-specific pipelines to implement FAIR-compliant repositories, and FAIR-compliant big data SRAs.
- A discussion encompassing the limitations and tendencies of the analyzed solutions, as well as the research opportunities that arise from these observations.

This paper is structured as follows. Section 2 describes the methodology and conduction, Section 3 presents the data synthesis and classification, Section 4 outlines limitations, tendencies, and research opportunities, and Section 5 concludes the paper.

2. Methodology and Conduction

A systematic review needs to follow a plan with phases and activities so that it can subsequently be reproduced. These are defined based in the work of Scannavino et al. (2017), as follows: (i) planning, encompassing the definition of an objective, research questions, search engines, keywords, search string, and selection criteria; (ii) conduction, including the studies selection and synthesis; and (iii) discussion of the results. This process is flexible in regards to reevaluating its phases, enabling their redefinition if deemed necessary.

For the planning phase, we first define the objective of the systematic review as "identifying studies that propose SRAs capable of implementing the FAIR Principles and addressing the intrinsic characteristics of big data environments". Then, from this objective, we derive research questions to verify if there are studies in the literature that propose: (i) general purpose big data SRAs; (ii) pipelines to implement specific FAIR-compliant repositories; or (iii) FAIR-compliant big data SRAs. Afterwards, we determine which search engines will be employed based on their reach, novelty, and availability for accessing the returned studies. Considering these criteria, we conduct our systematic review on IEEEXplore Digital Library[1], ACM Digital Library[2], and Elsevier Scopus[3].

To conduct the systematic review using these search engines, we need to derive the following artifacts from the previously defined research questions: (i) search keywords

[1]IEEEXplore Digital Library: `https://ieeexplore.ieee.org`
[2]ACM Digital Library: `https://dl.acm.org`
[3]Elsevier Scopus: `https://www.scopus.com`

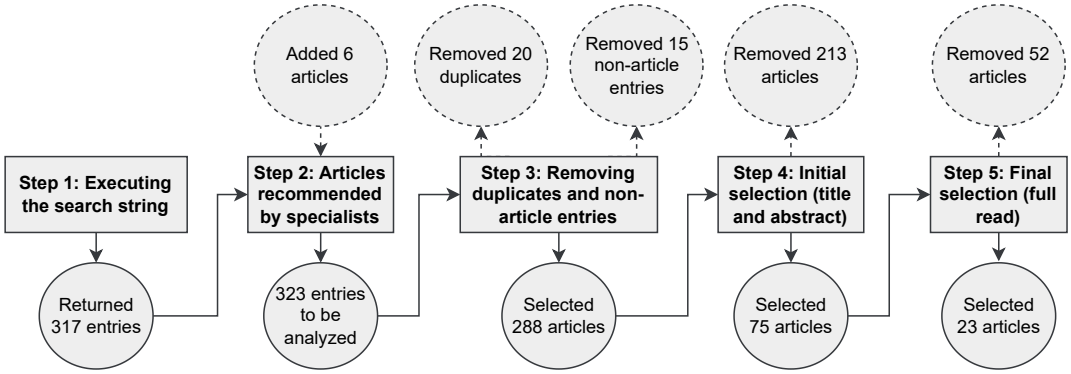


Figure 1. Conduction of the systematic review using the selection procedure.

and their synonyms; and (ii) a search string that connects these keywords using logical operators such as “AND" and “OR". We define the following search string, containing the derived keywords and the chosen synonyms:

```
(("FAIR principles" OR "FAIR guiding principles" OR "open science")
  AND ("implementation" OR "workflow" OR "pipeline"))
OR (("software reference architecture" OR "SRA" OR
     "generic architecture")
     AND (("big data" OR "cloud" OR "cloud computing")
           OR ("FAIR principles" OR "FAIR guiding principles" OR
               "open science")))
```

To delineate the selection criteria, we consider studies that: (i) address the research questions; (ii) propose architectural solutions; (iii) are freely accessible in academic environments; (iv) are available in English or Portuguese; (v) pertain to the field of Computer Science; and (vi) were published since 2020. To define this time frame, we leverage the studies of Jacobsen et al. (2020) and Van Reisen et al. (2020). These studies propose several recommendations for the implementation of the FAIR Principles based on a literature review conducted between the years of 2016, when the FAIR Principles were first proposed, and 2019. Since we are interested in solutions that adopt their recommendations during the definition of FAIR-compliant implementations, we disconsider studies that have been published prior to the year of 2020. The same time frame can be adopted for studies that define big data SRAs, since the work of Davoudian and Liu (2020) surveys the most relevant architectures in the literature up to the year of 2020. However, studies outside this time frame can also be included if they are recommended by specialists.

With the selection criteria defined, the systematic review can be conducted according to the selection procedure, as illustrated in Figure 1. In Step 1, we execute the search string in the chosen search engines, using their interface to limit the search to the article title, keywords, and abstract. We also use this interface to disconsider articles outside the scope of the selection criteria (iv), (v), and (vi), leaving only the criteria (i), (ii), and (iii) to be analyzed in subsequent steps. The search string retrieved a total of 317 entries[4], 11 from ACM, 43 from IEEEXplore, and 263 from Scopus. An equivalent search string with Portuguese keywords was also executed in the search engines, but no additional work was retrieved. Then, in Step 2, we include 6 references that have been recommended by

[4]We performed the search in April 8, 2023.

specialists, increasing the total number of entries to 323. However, 20 of these entries are duplicates and 15 do not correspond to scientific papers. We remove these in Step 3, leaving 288 studies to be analyzed, 8 from ACM, 41 from IEEEXplore, 233 from Scopus, and 6 from specialists recommendations.

Then, we perform an initial selection (Step 4) on the obtained studies by reading their title and abstract. If the study meets the selection criteria, it is approved to be analyzed in the next step. Of the 288 analyzed articles, 75 are approved in the initial selection, 2 of which are from ACM, 10 from IEEEXplore, 57 from Scopus, and 6 from specialists recommendations. The next step of the systematic review is the final selection (Step 5). In this activity, we read the studies selected in the previous step in their entirety to verify if they still meet the selection criteria. We approved 23 studies in this step, comprised of 4 from IEEEXplore, 13 from Scopus, and 6 from specialists recommendations. These studies are then synthesized and classified into groups, as detailed in Section 3.

3. Data Synthesis and Classification

We synthesize the content of the articles approved in the final selection and classify them into three distinct groups, based on which research questions they answer: (i) general purpose big data SRAs (Section 3.1); (ii) pipelines to implement specific FAIR repositories (Section 3.2); and (iii) FAIR-compliant big data SRAs (Section 3.3). We compare these studies based on their key characteristics. These encompass essential features for open science repositories (i.e. FAIR compliance, metadata management, source data retrieval by metadata), big data capabilities, storage of data and metadata, and being generic enough to adhere to the concept of an SRA. When relevant, we also analyze the repository context and the focus of the developed solutions.

3.1. Group 1: General Purpose Big Data SRAs

The research efforts allocated in this group encompass big data SRAs that have not been specifically engineered to align with the FAIR Principles (Table 1). Instead, these solutions emphasize the provision of real-time analytics to support users in decision-making, without focusing on the collection and management of metadata and data provenance.

The first work in this group describes the traditional data warehousing architecture [Chaudhuri and Dayal 1997]. It consists of a dedicated environment for the execution of analytical queries, encompassing components such as a data warehouse, data marts, and a metadata repository. The Kappa architecture [Kreps 2014] uses a single streaming layer for big data computation, supporting both batch and real-time processing through the buffering of historical data in a logging system for an extended duration. On the other hand, the Lambda architecture [Kiran et al. 2015] employs three layers for this type of computation, one for creating batch views, one for processing recent data into real-time views, and one to store and merge these views for later consumption.

Liquid [Fernandez et al. 2015] is an SRA that aims to overcome the limitations of Kappa and Lambda by employing incremental processing instead of fully recomputing views. It is comprised of two layers: (i) messaging, which stores data and metadata as messages, as well manages checkpoints from which this data can be partially recomputed; and (ii) processing layer, performing jobs and transformations on the messages. Although

Table 1. Comparison of general purpose big data SRAs.

Work*	Fits the concept of an SRA	FAIR-compliant	Metadata management	Source data retrieval by metadata	Enables big data analytics	Storage of data and metadata
Chaudhuri and Dayal (1997)	✓	✗	✓	✗	✓	Same infrastructure
Kreps (2014)	✓	✗	✗	✗	✓	Does not apply
Kiran et al. (2015)	✓	✗	✗	✗	✓	Does not apply
Fernandez et al. (2015)	✓	✗	✗	✗	✓	Does not apply
Martínez-Prieto et al. (2015)	✓	✗	✗	✗	✓	Does not apply
Nadal et al. (2017)	✓	✗	✓	✗	✓	Same infrastructure
Ataei and Litchfield (2021)	✓	✗	✓	✗	✓	Same infrastructure

*Architectures proposed prior to the year of 2020 are classified in this group due to their inclusion in the systematic review as recommendations from specialists.

the messaging layer stores metadata, no implementation regarding its management is defined on Liquid. The Solid architecture [Martínez-Prieto et al. 2015] unifies heterogeneous data into a single model, defining layers for big data storage, querying, streaming, and merging of historical and runtime data. It distinguishes itself from Lambda by storing data in a single layer instead of performing its duplication into batch and real-time views. As for the Bolster architecture [Nadal et al. 2017], the authors use a semantic layer with metadata management for data governance. This includes an ontology-based repository for input data characteristics, accompanied by a dispatcher component that determines whether the streaming data should be directed to the batch or speed layer.

Furthermore, NeoMycelia [Ataei and Litchfield 2021] is an SRA based on microservices and events. Each microservice has a local database with a caching component. This architecture is comprised of several components, such as: (i) a gateway for user connection; (ii) controllers and service meshes for stream and batch data processing; (iii) an event backbone and an event archive to support the communication between microservices; (iv) a data lake that stores structured, pseudo-structured, unstructured, and semi-structured data; (v) a query controller and query engine to support query execution; and (vi) a semantic layer, which contains a metadata management system responsible for storing metadata, preparation rules, and data evolution.

3.2. Group 2: Pipelines to Implement Specific FAIR Repositories

Studies classified in this group represent pipelines that implement the FAIR Principles to the context of specific data sharing repositories (Table 2). Solutions that do not implement data sharing repositories are disconsidered, such as FAIRness assessment tools or workflows to conduct specific scientific experiments.

In Assante et al. (2021), the authors propose the AGINFRA PLUS platform, which enables researchers to store, analyze, visualize, and publish agriculture and food data in accordance with open science. Moreover, Pană et al. (2021) developed a pipeline that extracts data and metadata from several seismic databases available online, integrating them into a single repository stored in a PostgreSQL local database to enable analytics. Furthermore, the work of Pestryakova et al. (2022) describes a pipeline that extracts data and metadata from COVID-19 related publications, transforming them into CovidPubGraph, a knowledge graph which, in itself, can be considered as a data sharing repository.

Table 2. Comparison of pipelines to implement specific FAIR repositories.

Work	Fits the concept of an SRA	FAIR-compliant	Metadata management	Source data retrieval by metadata	Enables big data analytics	Storage of data and metadata	Context
Assante et al. (2021)	✗	✓*	✓	✓	✗	Same infrastructure	Agriculture and food data
Pană et al. (2021)	✗	Partially	✓	✓	✗	Same infrastructure	Earthquake data
Pestryakova et al. (2022)	✗	✓	✓	✓	✗	Same infrastructure	COVID-19 papers data
Brůha et al. (2022)	✗	✓*	✓	✓	✗	Same infrastructure	Health and brain data
Jha et al. (2022)	✗	Partially	✓	✓	✗	Same infrastructure	Oncology data
Felikson et al. (2022)	✗	✓*	✓	✓	✓	Same infrastructure	Earth data from NASA
Borges et al. (2022)	✗	✓*	✓	✓	✗	Same infrastructure	COVID-19 patients data
Sciacca et al. (2022)	✗	✓*	✓	✓	✓	Same infrastructure	Underwater, atmospheric, and space data
Toulet et al. (2022)	✗	✓*	✓	✓	✗	Same infrastructure	Textual data from papers
Schwagereit et al. (2022)	✗	✓	✓	✓	✗	Same infrastructure	In vivo data
Deng et al. (2022)	✗	✓*	✓	✓	✗	Separate infrastructures	Immunology data
Rueda-Ruiz et al. (2022)	✗	✓*	✓	✓	✓	Same infrastructure	LiDAR data
Lehmann et al. (2023)	✗	✓*	✓	✓	✓	Same infrastructure	Sensor data

*The authors state that the pipeline is FAIR-compliant, however no details are given on how it fulfills each FAIR principle.

The Body in Numbers system [Brůha et al. 2022] encompasses the collection of health-related data and metadata taking into consideration the FAIR Principles. The authors propose a pipeline that consists of five modules to collect, annotate, analyze, interpret, and publish brain and physical data and its associated metadata. In Jha et al. (2022), the authors propose a pipeline to extract data and metadata from several healthcare systems. A series of Python scripts is employed to extract image features and perform data cleaning and integration, storing the result as data triples. Additionally, the work of Felikson et al. (2022) describes the cloud infrastructure behind the repository of NASA's Earth Information System. Its goal resides in enabling researchers and end users to conduct their own analyses close to big data stored in the cloud, and to make it easier to access data products and information, to reproduce analyses, and to build on existing work, following the concept of Open Science.

The VODAN BR project [Borges et al. 2022] aims to collect and implement a data management infrastructure for COVID-19 hospitalized patients' cases in Brazil, according to the FAIR principles. The authors describe its architecture, which covers the processes between the collection of clinical data to the publication of its metadata in the network as triplestores. Furthermore, NEANIAS [Sciacca et al. 2022] is a service-

oriented architecture to provide analytics for underwater, atmospheric and space related data. It is comprised of four high level core services to support open science lifecycles, integration with the open science cloud, artificial intelligence, and visualization. The work of Toulet et al. (2022) describes a pipeline to extract metadata from scientific papers and to generate knowledge from their full text, storing both as triplestores in a Virtuoso server. Moreover, in Schwagereit et al. (2022) the authors describe FISH, a platform to share in vivo data according to the FAIR Principles. It consists of multiple management and storage component available through microservices.

Deng et al. (2022) propose ImmuneData, a platform that extracts and integrates the metadata of different immunology databases into a single repository, which stores this metadata in a unified metadata model proper for biomedical data. This enables users to use a single engine to query this metadata to retrieve the source data objects. Additionally, in Rueda-Ruiz et al. (2022), the authors propose a general specification for cloud repositories to store large scale LiDAR data. It consists of a conceptual data model implemented on MongoDB and an API to handle requests. Finally, the work of Lehmann et al. (2023) proposes an architecture that merges the concepts of Research Data Management (RDM) based on the FAIR Principles with the concept of a digital twin, which is the virtual counterpart of a physical sensor. The architecture collects sensor data and metadata and sends them to a layer called RDM Core Space, where the data is stored in its raw format and the metadata is stored as a knowledge graph. Both the sensor data and metadata can then be used by smart applications through a messaging broker.

3.3. Group 3: FAIR-compliant Big Data SRAs

This group of studies encompasses a range of architectural frameworks that not only are generic enough to fit the concept of an SRA, but also are concerned with the requirements imposed by the FAIR Principles (Table 3). These studies also employ solutions to handle the intrinsic characteristics of big data environments (i.e. volume, variety, and velocity), such as parallel and distributed data processing and cloud computing technologies.

The work of Castro et al. (2022a) proposes BigFAIR, a FAIR-compliant SRA to store, process, and query scientific data and metadata. This architecture is comprised of several layers organized in two separate infrastructures: (i) local, encompassing the local environments of the data providers, where the source data objects remain stored; and (ii) repository, encompassing big data technologies for centralized metadata storage, data and metadata processing, and ad-hoc data anonymization. Metadata is stored either in the Metadata Lake in its raw format, or in the Metadata Warehouse after undergoing transformations, which ensures metadata persistence even when the associated source data objects no longer exist. The authors detail the compliance of each layer with each FAIR Principle. By taking advantage of the existing local infrastructures of data providers, this architecture is able to support data ownership and increase flexibility.

CloudFAIR [Castro et al. 2022b] is a FAIR-compliant SRA that handles both scientific data and its associated metadata in a single cloud infrastructure. The authors claim that this unification unburdens data providers in regards to the management of a local infrastructure and also improves performance. By being an extension of BigFAIR, CloudFAIR inherits its full compliance with the FAIR Principles, as well as the storage of transformed metadata in a Metadata Warehouse to guarantee its persistence. However,

Table 3. Comparison of FAIR-compliant big data SRAs.

Work	Fits the concept of an SRA	FAIR-compliant	Metadata management	Source data retrieval by metadata	Enables big data analytics	Storage of data and metadata	Focus
Castro et al. (2022a)	✓	✓	✓	✓	✓	Separate infrastructures	Flexibility, data ownership
Castro et al. (2022b)	✓	✓	✓	✓	✓	Same infrastructure	Performance, simplification for providers
Vazquez et al. (2022)	✓	Partially	✓	✓	✓	Same infrastructure	Metadata quality assurance

instead of using a Metadata Lake to store only metadata in its raw format, CloudFAIR uses a multi-tiered Data Lake that stores two copies of the source data objects: a fully anonymized copy and an encrypted copy without anonymization. The authors conduct a performance evaluation that proves that this storage strategy, along with the other features of CloudFAIR, improve performance in up to 75.95% when compared to BigFAIR.

Finally, the work of Vazquez et al. (2022) propose GADDS, a generic platform that stores research data and metadata in the cloud, achieving partial compliance with the FAIR Principles. Metadata is stored in a blockchain environment, which enforces metadata quality control since every entry is validated by every node in the network in a decentralized manner. A version control software is also employed to track changes in the metadata and guarantee its persistence even when the associated source data object is excluded. However, GADDS is unable to track changes in the data objects. Thus, if a data object changes, the older versions of its associated metadata will point to the novel version of the data object. These data objects are stored by GADDS in an object storage in the cloud that enables data replication into multiple nodes, allowing parallel and distributed data processing. The authors validate GADDS with a case study related to tissue engineering and discuss its FAIR compliance at a high level.

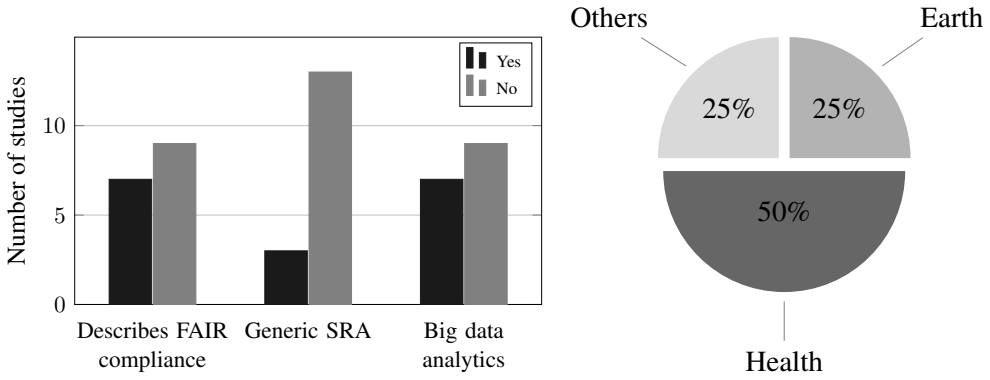
4. Discussion

4.1. Limitations

The objective of this systematic review resides in retrieving studies that propose SRAs capable of implementing the FAIR Principles while also addressing the intrinsic characteristics of big data environments. However, the studies synthesized in Section 3 face some limitations in this regard, described as follows and depicted in Figure 2a.

Studies in Group 1 (Section 3.1) inherently diverge from the concept of a FAIR-compliant SRA in regards to their purpose. First, these SRAs fall short in meeting the requirements set by the FAIR Principles due to their limited capabilities for retrieving source data objects based on metadata and for keeping metadata alive even when the associated data objects are no longer available. Additionally, these architectures do not employ a specific component for storing metadata about the metadata. They either do not store this content or store all the metadata in the same component, compromising the richness and performance of metadata analyses.

Regarding Group 2 (Section 3.2), none of the reviewed studies propose architectures generic enough to fit the concept of an SRA. Rather, they propose pipelines that



(a) Main limitations. **(b)** Data contexts.

Figure 2. Main limitations and data contexts of studies in Groups 2 and 3.

are specific to implement a FAIR-compliant repository in a given context. Furthermore, the majority of these studies fail to clarify which of the FAIR Principles are satisfied by their solutions. Only the studies of Pestryakovaet al. (2022) and Schwagereit et al. (2022) detail how their full FAIR compliance is achieved, while the studies of Pană et al. (2022) and Jha al. (2022) clarify their partial compliance. This lack of information gives rise to substantial concerns, such as making sure that all principles are fulfilled and identifying which part of the solution is responsible for implementing each principle. Studies in this group are also limited in regards to their big data capabilities. Only the studies of Felikson et al. (2022), Sciacca et al. (2022), Rueda-Ruiz et al. (2022), and Lehmann et al. (2023) employ big data technologies in the construction of their solutions. Although enabling big data analytics is not a requirement imposed by the FAIR Principles, it is of significant importance to support the decision-making process. It not only allows data consumers to perform different types of analyses on the stored data and metadata, but also contributes to an increase in the overall performance of the repository.

The studies classified in Group 3 (Section 3.3) also present some limitations. For instance, GADDS [Vazquez et al. 2022] is unable to achieve full compliance with the FAIR Principles. By storing metadata in a blockchain environment, it can only be exposed to members inside the network. This hinders the general public unable to access the content of a repository implemented by GADDS, compromising findability and accessibility. Also, this SRA does not implement global unique identifiers, further impacting on its FAIR compliance. Another limitation of the studies in this group is related to their focus. For instance, CloudFAIR [Castro et al. 2022b] uses a single cloud infrastructure to store data and metadata to improve performance and to unburden data providers. However, by doing so it relinquishes support to data ownership and flexibility, key features of BigFAIR [Castro et al. 2022a]. This can be a problem in situations in which the repository is required to comply with different data protection regulations, or to implement specific security policies. The reverse situation is also true: BigFAIR forsakes performance and simplification for data providers in order to support data ownership and obtain flexibility. Furthermore, by not using a blockchain environment like GADDS, both BigFAIR and CloudFAIR relinquish decentralized metadata control, negatively impacting on the quality assurance of stored metadata.

4.2. Tendencies

After analyzing the data synthesis obtained in this systematic review (Section 3), we can identify some tendencies in the described studies. For instance, applying the FAIR Principles to health-related data is a common occurrence. Between the studies of Group 2 (Section 3.2), the following adhere to this context: (i) Pestryakova et al. (2022), with data from COVID-19 research publications; (ii) Brůha et al. (2022), using health and brain data; (iii) Jha et al. (2022), employing oncology related data; (iv) Borges et al. (2022), using data from COVID-19 patients; and (v) Deng et al. (2022), leveraging immunology data. Additionally, both BigFAIR [Castro et al. 2022a] and CloudFAIR [Castro et al. 2022b] use COVID-19 patients data in their experiments, whereas GADDS [Vazquez et al. 2022] employs fiber cell tissue research data during its instantiation. Another commonly addressed context is that of earth-related data, being covered by the repositories of Pană et al. (2021), Felikson et al. (2022), Sciacca et al. (2022), and Rueda-Ruiz et al. (2022). The proportion of studies per data context is illustrated in Figure 2b.

Another observed tendency is the adoption of microservices for the construction of architectures and pipelines, which is a paradigm to deploy applications as a collection of events that are inherently independent. This strategy is observed in NeoMycelia [Ataei and Litchfield 2021], the latest general purpose big data SRA available in the literature, and in the FISH platform [Schwagereit et al. 2022], a pipeline to implement an in vivo data repository in accordance with the FAIR Principles. Finally, the employment of the same infrastructure for the management of data and metadata is another detected tendency. The majority of the surveyed solutions use this strategy, with only a few exeptions [Deng et al. 2022, Castro et al. 2022a]. According to the experiments conducted in Castro et al. (2022b), a possible reason for the occurrence of this tendency is the improvement of query performance when storing data and metadata in the same infrastructure. This strategy also unburdens data providers in regards to maintaining a local repository to store scientific data, while overlooking data ownership and flexibility.

4.3. Research Opportunities

The aforementioned limitations and tendencies give rise to opportunities to conduct innovative research. For instance, the development of a novel FAIR-compliant SRA capable of unifying the advantages of BigFAIR [Castro et al. 2022a], CloudFAIR [Castro et al. 2022b], and GADDS [Vazquez et al. 2022] would considerably benefit the scientific community. However, leveraging flexibility, data ownership, performance, simplification for data providers, and metadata quality assurance in a single architecture is challenging. A possible solution is developing this SRA in multiple modules, each prioritizing one of the aforementioned characteristics. These modules can then be instantiated depending on the requirements imposed by the data providers and consumers.

Another opportunity that arises from the previously identified tendencies is the development of a FAIR-compliant SRA using the strategy of microservices. This strategy has successfully been employed by Ataei and Litchfield (2021) for developing a general purpose big data SRA and by the work of Schwagereit et al. (2022) for implementing a FAIR-compliant repository. However, we have not identified a solution that merges these research fields during the conduction of the systematic review, representing a gap that can be further explored by the scientific community.

5. Conclusions and Future Work

In this paper, we presented a systematic review of the literature that identified architectural solutions capable of implementing the FAIR Principles and addressing the intrinsic characteristics of big data environments. We detailed its methodology and conduction, enabling reproducibility. Moreover, we introduced a data synthesis of the selected studies, as well as their classification in three distinct groups. We also identified the limitations of these solutions, deriving tendencies and research opportunities. Future work consists on analyzing a broader scope of studies by considering the snowball technique. We also plan on providing future updates for this systematic review and to explore the identified research opportunities, proposing novel architectural solutions for the implementation of the FAIR Principles in big data environments.

Acknowledgments

This work was supported by the São Paulo Research Foundation (FAPESP), the Brazilian Federal Research Agency (CNPq), and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil (CAPES), Finance Code 001. Cristina D. Aguiar was supported by the grant #2018/22277-8 (FAPESP).

References

Assante, M. et al. (2021). Realising a science gateway for the agri-food: the AGINFRA PLUS experience. In *CEUR Workshop Proc.*

Ataei, P. and Litchfield, A. (2021). NeoMycelia: A software reference architecture for big data systems. In *Proc. APSEC*, pages 452–462.

Borges, V. et al. (2022). A platform to generate FAIR data for COVID-19 clinical research in Brazil. In *Proc. ICEIS*, pages 218–225.

Brůha, P. et al. (2022). Workflow for health-related and brain data lifecycle. *Front. Digit. Health*, 4.

Castro, J. P. C. et al. (2022a). FAIR Principles and Big Data: A software reference architecture for Open Science. In *Proc. ICEIS*, pages 27–38.

Castro, J. P. C. et al. (2022b). Open Science in the cloud: The CloudFAIR architecture for FAIR-compliant repositories. In *Proc. ADBIS*, pages 56–66.

Chaudhuri, S. and Dayal, U. (1997). An overview of data warehousing and OLAP technology. *SIGMOD Rec.*, 26(1):65–74.

Chen, M., Mao, S., and Liu, Y. (2014). Big data: A survey. *Mob. Netw. Appl.*, 19(2):171–209.

Davoudian, A. and Liu, M. (2020). Big data systems: A software engineering perspective. *ACM Comput. Surv.*, 53(5):1–39.

Deng, N. et al. (2022). ImmuneData: an integrated data discovery system for immunology data repositories. *Database*, 2022.

Felikson, D. et al. (2022). NASA's earth information system: Sea-level change. In *OCEANS 2022, Hampton Roads*, pages 1–8.

Fernandez, R. C. et al. (2015). Liquid: Unifying nearline and offline big data integration. In *Proc. CIDR*.

Jacobsen, A. et al. (2020). FAIR principles: interpretations and implementation considerations. *Data Intell.*, 2(1-2):10–29.

Jha, A. K. et al. (2022). Implementation of big imaging data pipeline adhering to FAIR principles for federated machine learning in oncology. *IEEE Trans. Radiat. Plasma Med. Sci.*, 6(2):207–213.

Kiran, M. et al. (2015). Lambda architecture for cost-effective batch and speed big data processing. In *IEEE Trans. Big Data*, pages 2785–2792.

Kreps, J. (2014). Questioning the Lambda architecture. Available at `https://www.oreilly.com/radar/questioning-the-lambda-architecture/`. Accessed in April 8, 2023.

Lehmann, J. et al. (2023). Establishing reliable research data management by integrating measurement devices utilizing intelligent digital twins. *Sensors*, 23(1):468.

Martínez-Prieto, M. A. et al. (2015). The solid architecture for real-time management of big semantic data. *Future Gener. Comput. Syst.*, 47:62–79.

Medeiros, C. B. et al. (2020). *IAP input into the UNESCO Open Science Recommendation*. Available at `https://www.interacademies.org/sites/default/files/2020-07/Open_Science_0.pdf`. Accessed in April 8, 2023.

Nadal, S. et al. (2017). A software reference architecture for semantic-aware big data systems. *Inf. Softw. Technol.*, 90:75–92.

Nakagawa, E. Y., Antonino, P. O., and Becker, M. (2011). Reference architecture and product line architecture: A subtle but critical difference. In *Proc. ECSA*, pages 207–211.

Pană, G. T. et al. (2021). Towards the implementation of FAIR principles on an earthquake analysis platform. In *Proc. RoEduNet*, pages 1–4.

Pestryakova, S. et al. (2022). CovidPubGraph: A FAIR knowledge graph of COVID-19 publications. *Sci. Data*, 9(1):389.

Rueda-Ruiz, A. J. et al. (2022). SPSLiDAR: towards a multi-purpose repository for large scale LiDAR datasets. *Int. J. Geogr. Inf. Sci.*, 36(5):992–1011.

Scannavino, K. R. F. et al. (2017). *Revisão Sistemática da Literatura em Engenharia de Software: Teoria e Prática*. Elsevier.

Schwagereit, F. et al. (2022). FAIR data APIs in the FAIR in vivo data sharing platform. In *CEUR Workshop Proc*.

Sciacca, E. et al. (2022). Scientific visualization on the cloud: the NEANIAS services towards EOSC integration. *J. Grid Comput.*, 20(1):7.

Toulet, A. et al. (2022). ISSA: generic pipeline, knowledge model and visualization tools to help scientists search and make sense of a scientific archive. In *Proc. ISWC*, pages 660–677.

Van Reisen, M. et al. (2020). Towards the tipping point for FAIR implementation. *Data Intell.*, 2(1-2):264–275.

Vazquez, P. et al. (2022). Globally accessible distributed data sharing (GADDS): A decentralized FAIR platform to facilitate data sharing in the life sciences. *Bioinformatics*, 38:3812–3817.

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, 3(1):1–9.