An examination of scientific data repositories, data reusability, and the incorporation of FAIR

Angela P. Murillo

Abstract
Scientific data repositories (SDRs) provide a way for scientists to share data through data deposition and reuse of deposited data. Over the last twenty-plus years, hundreds of scientific SDRs have become available. This research examines 132 SDRs. This study assesses if the information available in the SDRs aligned with what scientists need to determine data reusability and if the SDRs enforce FAIR principles.

KEYWORDS
data reusability, FAIR principles, scientific data repositories

1 | INTRODUCTION

The creation of scientific data repositories (SDRs) has provided scientists the ability to share and reuse data by making data readily available through online websites where scientists can search for data to reuse. However, data availability is only one important factor for data sharing and reuse. As discussed in previous research, it is important to consider the nature of shareability, availability, and reusability, all of which impact the data reuse process (Yoon, Jeng, Curty, & Murillo, 2017).

While SDRs have increased the availability of data, the information provided on the data record is vital to determine if the data is truly reusable. Previous research has described how a data record can be defined as structured information that presents essential information about a data product (Liu, Grossman, & Zhai, 2003). Additionally, previous research has discussed how, when assessing data reusability, important to consider if the data are relevant, and if the data can be understood (Faniel & Jacobsen, 2010). This information can only be determined through the data record. Furthermore, previous research has examined data records to determine what information scientists need to determine data reusability. This research also included recommendations for a data record prototype (Table 1) of what is considered essential information for data reusability (Murillo, 2019). Along with the recommended information for data records, FAIR principles have provided further guidance for findability, accessibility, interoperability, and reusability of data (Table 2) (FORCE11, 2014).

This poster presents a preliminary analysis of ongoing research that examines if SDRs adhere to best practices for providing the necessary information to determine reusability and if they adhere to FAIR principles.

2 | METHODS

The research questions of this study were to determine if the information available in scientific data repositories (SDRs) aligns with scientist’s reusability needs. We used two metrics to analyze this question. First, we considered if the information available aligned with a model of best practices for data reusability (Murillo, 2019), and additionally, we examined if the SDRs aligned with the FAIR principles (FORCE11, 2014).
There are hundreds of SDRs we could have considered. However, we decided to scope this research by reviewing the 100 SDRs listed from previous research (Marcial & Hemminger, 2010) and the SDRs listed on the PLoS Recommended Repositories webpage (PLOS ONE, 2020).

A total of 132 SDRs were reviewed.

We conducted a content analysis of the SDRs. During the initial analysis, we determined the discipline of the SDR, if the SDR still existed, and if the SDR imposed FAIR standards. We then examined three data records from each active SDR for a total of 378 records to determine if the data records aligned with what is considered essential information for data records from the previous research. Both researchers reviewed each data record to ensure agreement on all items examined.

### Results

Of the 132 SDRs examined, 126 were still active, and six were either closed or inactive. Of the 126 active SDRs

---

**TABLE 1** Data record prototype: ideal attributes and definitions (Murillo, 2019)

| Attribute information | Attribute definition |
|-----------------------|----------------------|
| Data Description      | Short and succinct data description. |
| Data Creator          | Who collected the data? |
| Data Format           | What is the format of the data? (.csv, .txt, .tif, etc.) |
| Data Type             | What type of data was collected? (field, experiment, sensor, simulation) |
| Data Size             | What is the size of the data? (MBs, TBs, etc.) |
| Data Collection Location | Where was the data collected? |
| Data Date Range       | When was the data collected? |
| Research Methods Information | How was the data collected, by what means, what were the steps involved? |
| Instrument Information | What instruments were used to collect the data and what were the calibration settings? |
| Provenance Information | Was the data changed in any way, if so how, why, and by what methods and/or instrument? |
| Data Abstract         | A descriptive summary of the data. The data abstract should describe the data, not the paper associated with the data. |
| Attribute and Unit Lists | This includes data variables and how these were measured. |

**Secondary information**

| Attribute | Attribute Definition |
|-----------|----------------------|
| Taxonomic Information | If appropriate for the dataset, any information regarding biological organisms. |
| Data Citation and Persistent Identifier | A suggested data citation format, and DOI or other persistent identifier for the data. |
| Intellectual Rights Information | Any statement regarding restrictions to use of the data, as well as attribution instructions. |
| Data Keywords | Keywords linked throughout repository so that potential reusers can click to similar datasets. |
| Metadata Standard | Metadata standard used by the data. |
| Funding Source | Funding source for the data collection. |
| Publication Date | Date the data was published. |

---

**TABLE 2** FAIR Data Principles (FORCE11, 2014)

| The Fair guiding principles |
|-------------------------------|
| **Findable** |
| F1. (Meta)data are assigned a globally unique and persistent identifier |
| F2. Data are described with rich metadata |
| F3. Metadata clearly and explicitly include the identifier of the data they describe |
| F4. (Meta)data are registered or indexed in a searchable resource |
| **Accessible** |
| A1. (Meta)data are retrievable by their identifier using a standardized communications protocol |
| A1.1 The protocol is open, free, and universally implementable |
| A1.2 the protocol allows for an authentication and authorization procedure, where possible |
| A2 Metadata are accessible, even when the data are no longer available |
| **Interoperable** |
| I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation |
| I2. (Meta)data use vocabularies that follow FAIR principles |
| I3. (Meta)data include qualified reference to other (meta)data |
| **Reusable** |
| R1. Meta(data) are richly described with a plurality of accurate and relevant attributes |
| R1.1 (Meta)data are released with a clear and accessible data usage license |
| R1.2 (Meta)data are associated with detailed provenance |
| R1.3 (Meta)data meet domain-relevant community standards |
examined, 46 (35%) explicitly stated that they were adhering to FAIR principles. Seventeen disciplines were represented in the SDRs, including earth and environmental sciences, biology, astronomy, social sciences, medicine, and astronomy.

Regarding essential information for data reusability, the researchers reviewed if the data records included the following information: Creator, Format, Type, Location, Date Range, Research Methods, Instrument Information, Provenance Information, Data Abstract, Attribute/Unit List, and Data Description.

Of the 126 active SDRs, 42 (32%) provided the majority of the essential information for data reusability.

Table 3 provides a summary of how often essential information was provided of the 42 records that included the essential information as part of the data record.

Additionally, essential information that can contain nuance, such as research methods, provenance information, and data abstract, were analyzed for robustness. It was determined that the majority of the records that did provide this information included quite robust information.

| Essential information         | Percentage of records |
|------------------------------|-----------------------|
| Data creator                 | 95%                   |
| Data format                  | 81%                   |
| Data description             | 75%                   |
| Data collection date         | 67%                   |
| Data collection location     | 65%                   |
| Data type                    | 60%                   |
| Attribute/unit list          | 53%                   |
| Instrument information       | 53%                   |
| Data abstract                | 53%                   |
| Research methods             | 46%                   |
| Data size                    | 43%                   |
| Provenance information       | 42%                   |

4 | CONCLUSION

This research provides an analysis of 132 scientific data repositories to determine if SDRs are adhering to best practices for data reusability and FAIR principles. While many SDRs exist, data within these repositories only become useful to scientists if they can determine reusability through the information made available in the data record. FAIR principles provide a way to assist in ensuring data reusability.

A continued analysis is being conducted on this data. This future work will include an analysis of how SDRs are adhering to FAIR principles, an analysis of repository recommendations from FAIRsFAIR (FAIRsFAIR, 2020), a more in-depth analysis of the SDR data records, and a review of SDR disciplinary differences. This research aims to assist in determining which SDRs are adhering to best practices to assist researchers in determining where best to deposit data.

REFERENCES

FAIRsFAIR (2020). FAIRsFAIR: Fostering fair data practices in Europe. FAIRsFAIR. https://www.fairsfair.eu/

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. Computer Supported Cooperative Work (CSCW), 19(3), 355–375. https://doi.org/10.1007/s10606-010-9117-8

FORCE11. (2014, September 10). Guiding principles for findable, accessible, interoperable and re-usable data publishing version b1.0. FORCE11. Retrieved from https://www.force11.org/fairprinciples

Liu, B., Grossman, R., & Zhai, Y. (2003). Mining data records in web pages. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 601–606. https://doi.org/10.1145/956750.956826

Marcial, L. H., & Hemminger, B. M. (2010). Scientific data repositories on the Web: An initial survey. Journal of the American Society for Information Science and Technology, 61(10), 2029–2048. https://doi.org/10.1002/asi.21339

Murillo, A. P. (2019). Data matters: How earth and environmental scientists determine data relevance and reusability. Collection and Curation. https://doi.org/10.1108/CC-11-2018-0023.

PLOS ONE. (2020). PLOS One. Recommended Repositories. Retrieved from https://journals.plos.org/plosone/s/recommended-repositories

Yoon, A., Jeng, W., Curty, R., & Murillo, A. (2017). In between data sharing and reuse: Shareability, availability and reusability in diverse contexts. Proceedings of the Association for Information Science and Technology, 54(1), 606–609. https://doi.org/10.1002/pra2.2017.14505401085

How to cite this article: Murillo AP. An examination of scientific data repositories, data reusability, and the incorporation of FAIR. Proc Assoc Inf Sci Technol. 2020;57:e386. https://doi.org/10.1002/pra2.386