Automatic FAIR Provenance Collection and Visualization for Time Series

Fadoua Rafii
Digital Technology Skills Ltd
Dublin, Ireland
fadoua.rafii@ncirl.ie

Horacio González–Vélez
National College of Ireland
Dublin, Ireland
horacio@ncirl.ie

Adriana E. Chis
National College of Ireland
Dublin, Ireland
adriana.chis@ncirl.ie

ABSTRACT

Provenance provides data lineage and history of different transformations applied to a dataset. A complete trace of data provenance can enable the reanalysis, reproducibility, and reusability of features, which are essential for validating results and extending them in many projects. Open time series datasets are readily accessible and discoverable, but their full reproducibility and reusability require clear metadata provenance. This paper introduces an assessment of provenance variables using an algorithm for collecting FAIR (Findable, Accessible, Interoperable, Reusable) characteristics in open time series and generating an associated provenance graph. We have evaluated the FAIRness of provenance traces by automatically mapping their properties to a provenance data model graph for a case study employing open time series from weather stations. Our approach arguably enables researchers to analyse time series datasets with similar characteristics, prompting new research questions, insights, and investigations. As a result, this approach has the potential to promote reusability and reproducibility, which are critical factors in scientific research.

CCS CONCEPTS

• **Information systems** → **Digital libraries and archives**; • **Applied computing** → Annotation; *Document metadata.*

KEYWORDS

Data Provenance, FAIR, Time series, Reproducibility, Reusability, Metadata

ACM Reference Format:
Fadoua Rafii, Horacio González–Vélez, and Adriana E. Chis. 2023. Automatic FAIR Provenance Collection and Visualization for Time Series. In *Companion of the 2023 ACM/SPEC International Conference on Performance Engineering (ICPE '23 Companion), April 15–19, 2023, Coimbra, Portugal.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3578245.3585026

1 INTRODUCTION

In 2016, a group of stakeholders formalised a measurable set of guidelines to encourage the knowledge discovery and increase the datasets utility. Known as the FAIR (Findable, Accessible, Interoperable, Reusable) principles [29], they have become increasingly crucial to enable researchers and practitioners to access, find, understand, and process datasets. However, the actual operationalisation of those principles is challenging and needs guidance [18], as experience has shown that several conditions need to be fullfiled to completely enable data sharing, reusability, and reproducibility [28].

Typically used as a guide to authenticity or quality, *provenance* has been applied to scientific workflows for reproducibility, i.e. to keep a detailed record of the steps to produce a result, ultimately simplifying exploratory processes, fostering collaboration, and enabling knowledge transfer [8]. In fact, the provenance of scientific results—including how results were obtained, what datasets were used as input, and what parameters have an impact on the derivation—has been deemed crucial to reproduce the whole scientific process [12], and formal educational programmes have started to emerge to train scientists and practitioners in FAIR principles with emphasis on open science and research data management [13].

The usefulness of weather data is provided by its information and metadata that can be exploited in multiple sectors, domains, applications, and studies [19]. Weather variables are utilised for forecasting in smart grid systems [2]. Features like precipitation, air temperature, humidity and wind speed give insight into several energy-related forecasting problems due to the connection of energy and weather demand [19].

Furthermore, weather conditions represent an important factor that might profoundly have an impact on daily life [1]. Walmart used weather data for marketing decisions, and they acted on the correlations that exist between weather and store sales to merchandise the store-level and advertise the hyper-local digital [6]. Several studies have shown that weather influences people's mood states, social behaviour and trading decisions [9, 17, 20].

Due to the utility of weather data, provenance is arguably critical to ensuring the comprehensive history of processes that are applied to the features from their origin to their current state. In this paper, we evaluate the compliance of weather provenance metadata with FAIR provenance definition based on Research Data Alliance (RDA) and GO-FAIR recommendations.

We propose a systematic way of retrieving provenance properties and converting the event traces into a graph using the W3C PROV Data Model [24]. Our approach arguably allows climate researchers and scientists to clearly visualize data variations between several weather stations records without repetitively accessing resources/sensors and retrieving the distributed metadata manually for each time series.

PFAIR '23, April 15–16,2023, Coimbra, Portugal

ACM ISBN 979-8-4007-0072-9/23/04…$15.00
https://doi.org/10.1145/3578245.3585026

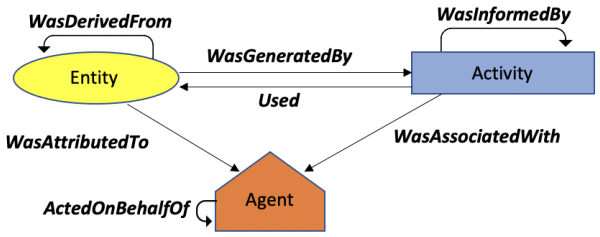

Figure 1: Components of the W3C PROV Data Model (PROV-DM). Adapted from [23].

2 RELATED WORK

The initial FAIR stakeholders established that a number of public databases applied standards to provide rigorous formal documentation of each dataset and its provenance with the aim of ensuring both interoperability and reusability [14]. However, it is not always the case for all general-purpose datasets. For this reason, reusability is challenging especially for undocumented data. Within the scientific research community, this has generated a reproducibility crisis [3]. For better understanding whether data can be reused, it is critical to understand how data fits in a larger context [15]. It comprises the understanding of how data were generated, their provenance [16], relationships to other sources and unambiguous concepts description [15].

The recommended W3C PROV Data Model (PROV-DM) defines the *Type* triad comprising an *entity* representing a digital or physical thing; an *activity* encompassing an action to use or create entities; and, an *agent* in charge of an activity [23]. Within PROV-DM, the 7 provenance relations are indicated by labelled edges to express Generation, Usage, Communication, Derivation, Attribution, Association, and Delegation. In Figure 1, we have represented an entity with a yellow oval, an activity with a blue rectangle, and an agent with an orange pentagon. Thus, we will use the two main PROV-DM concepts:

(1) Types: Entities, Activities, and Agents
(2) Relations: *wasGeneratedBy*, *used*, *wasInformedBy*, *wasDerivedFrom*, *wasAttributedTo*, *wasAssociatedWith*, and *actedOnBehalfOf*.

A novel approach to capture, manage, and publish the provenance metadata collected in environmental monitoring workflows based on PROV-DM triad has been previously discussed in the literature [25]. Their directed graph shows the lifecycle of a dataset generated by a hypothetical process of environmental sensor-based monitoring, and their raw dataset entity is related to capture activity through a `WasGeneratedBy` relationship. The capturing workflow uses electronic equipment and functional sittings, and it is associated with the institution to which the dataset is attributed. Such approach aims to manage the provenance records of environmental monitoring processes using an ad-hoc computational architecture for heterogeneous and distributed environments. The validation of this approach is assessed by implementing the architecture to manage the generated provenance metadata during the execution of a simulation. The provenance services allow to query the provenance metadata of the produced data, its generation workflow, and other relevant records. Their results shed light on the effectiveness of their approach when collecting, storing and querying the provenance of products metadata. Furthermore, it enables visualization and exploration of raw data, scientists involved and processes.

Table 1: Comparison of provenance implementations

Approach	Related-readings	PROV-DM	PROV-Template
[7]	Environmental monitoring	✓	X
[25]	Sensors	✓	X
Our proposal	Stations	✓	✓

Da Silva et al. [7] introduce a provenance-sensor model and explained how sensor readings could be converted to provenance. They have used graphic representations for PROV-DM as well. Their graph depicts data dependency between device and three sensors, and sensors are connected to events. Based on the developed model, they illustrate the case of a device with one sensor of humidity and temperature, the event is defined by $e = temperature, humidity$. They achieved the goal of introducing provenance awareness for Internet of Things (IoT) System through a framework of provenance collection for IoT devices. The effectiveness of their framework is evaluated through a prototype system that demonstrates that the framework can automatically collect provenance records in an IoT system.

Stoffers et al. [27] have developed the Backbone Catalogue of Relational Debris Information (BACARDI) to work on the classification of space objects orbiting the Earth. Designed to track distinct workflows, the BACARDI core components collect provenance at task level via manual developer annotation and predefined expressions. While effective, manual annotations can become error prone, particularly for large complex workflows.

2.1 Contribution

The FAIR Data Principles dictate the types of behaviours that data stewards and researchers can expect from digital resources [21], and impose the compliance-requirements on researchers if they want to publish their outputs FAIRly [30]. The application of the FAIR principles depends critically on rich metadata, yet domain vocabularies are mostly underused [26]. In this exploratory study, our aim is to investigate some randomly selected weather time series to check FAIRness provenance-compliance of metadata.

As part of Smardy–an EU-funded project which is deploying a traceable FAIR-compliant open innovation marketplace for data [11]–this paper aims to enhance weather datasets exploitation through provenance information. It presents a description of entities and processes involved in generating these datasets. Metadata of workflow provenance is generated based on the W3C PROV data model. Moreover, the paper provides an overview of the steps followed to automatically generate provenance records using as use case datasets from the Irish meteorological service Met Éireann.

Figure 2 summarises our approach. By extending the BACARDI provenance primitives via systematic annotation and collection, we use Met Éireann time series as input to generate a provenance data model graph with a clear view of the provenance trace. First,

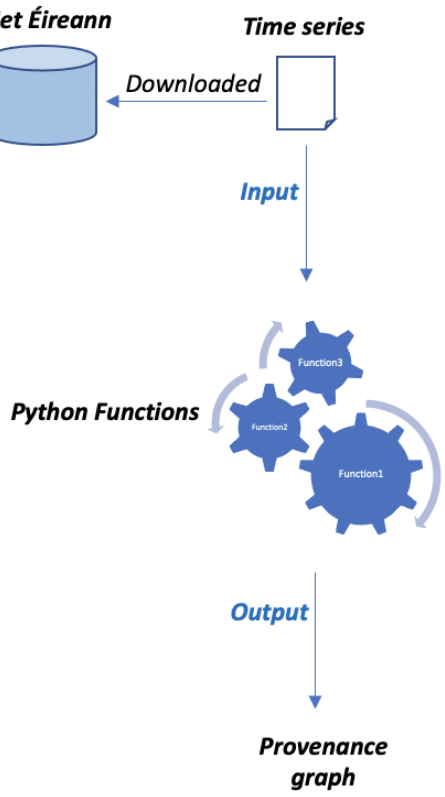

Figure 2: Workflow of provenance records collection and visualization. (Examples of provenance graphs are presented later in Fig. 3 and Fig. 4.)

Table 2: Selected time series datasets

Dataset	Location	Shape
1	Belmullet	(580354, 21)
2	Cork airport	(533999, 21)

we automatically extract the provenance properties responding to FAIR-related questions for building the provenance tracking. Then, we feed forward the resulting records to the PROV-DM graph components. These two steps have been executed through BACARDI's Python primitives deployed to automatically collect and visualize metadata.

3 MATERIALS AND METHODS

According to the RDA, a provenance is determined as "an indicator that requires the metadata to include information about the provenance of the data, i.e., information about the origin, history or workflow that generated the data, in a way that is compliant with the standards that are used in the community for which the data is curated" [10]. On the other hand, GO FAIR defines provenance as knowledge of "where the data came from (i.e., the clear story of origin/history), who to cite and/or how you wish to be acknowledged." [1].

In our study, provenance enables us to answer key question such as *"Where did a particular datum come from?"*, i.e. the exact geographical coordinates of the generating station which is a key metadata field, as accurate latitude and longitude for different stations can significantly impact climate models.

For provenance record generation, we have employed the PROV-Template [22] to inform the topology of the weather time series produced by each provenance instance. A template has variables for values, and these values are logged as a data structure binding that associates values and variables [4].

Listing 1: Serialization of the provenance graph shown in Fig. 3

```
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix ex: <http://example/> .
@prefix geo: <http://www.w3.org/2003/01/geo/
    wgs84_pos#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema
    #> .

{
    ex:Weather_DataSet a prov:Entity ;
        dcterms:created "2022-12-07T15:42:40"
            ^^xsd:dateTime ;
        dcterms:format "csv"^^xsd:string ;
        dcterms:modified "2023-01-23T14:28:10"
            ^^xsd:string ;
        dcterms:source "https://www.met.ie/
            climate/available-data/historical-
            data"^^xsd:string ;
        dcterms:title "hly2375"^^xsd:string ;
        prov:wasGeneratedBy ex:Create .

    ex:Create a prov:Activity ;
        prov:endedAtTime "2022-01-12T00:00:00"
            ^^xsd:dateTime ;
        prov:startedAtTime "1956-09-16T15
            :00:00"^^xsd:dateTime ;
        prov:wasAssociatedWith ex:Station .

    ex:Station a prov:Agent ;
        geo:elev " 9 M "^^xsd:string ;
        geo:lat "54.228  "^^xsd:string ;
        geo:lon " -10.007"^^xsd:string ;
        prov:atLocation " BELMULLET"^^xsd:
            string .
}
```

Data from the Irish meteorological service, Met Éireann, was selected. Two weather observing stations from Ireland are randomly chosen. The first selected dataset consists of hourly real-life weather features from 1956, and the second one contains hourly records from 1962.

The datasets locations and their shapes are described in Table 2. The time series contains records of all the available features such

[1]https://www.go-fair.org/fair-principles/r1-2-metadata-associated-detailed-provenance/

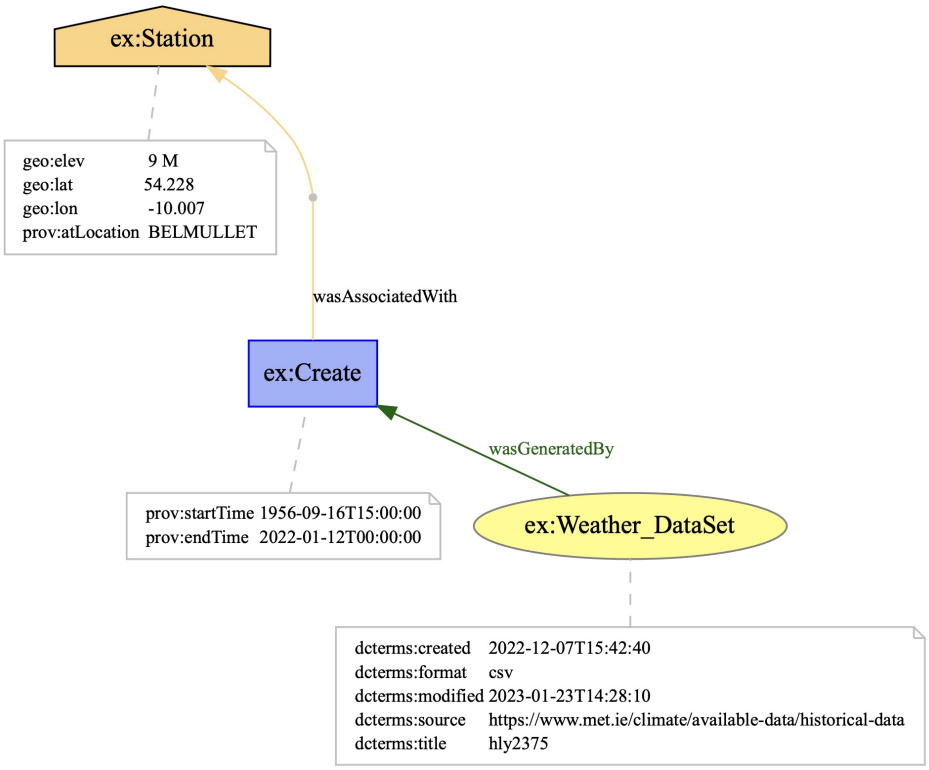

Figure 3: Provenance of Belmullet station records.

as air temperature, precipitation, relative humidity and vapour pressure. These datasets are used to generate and test an appropriate model to collect provenance records based on the FAIR guidelines.

4 IMPLEMENTATION AND RESULTS

Based on the two Provenance data models [7, 25], Figure 3 and Figure 4 show two provenance templates, the connection between the weather dataset (`ex: Weather_DataSet`) and the creation activity (`ex:Create`) is referred by `WasGeneratedBy` relationship. The creation workflow is associated with the weather station entity (`ex:Station`). The weather dataset entity holds the information about the dataset, such as its creation date, modification date, file format, name and source. The station entity contains information about the station's location, longitude, latitude, and height. The creation activity has information about the start time and the end time (`prov:startTime`, `prov:endTime`), that shows the selected time interval of the dataset. The generated provenance graph can be serialized by using the Resource Description Framework (RDF) [5], as depicted in Listing 1.

We have demonstrated how the use of provenance graphs can support reusability and reproducibility, using real station records from the publicly accessible repository of "Met Éireann". Our approach, which utilises the PROV family for provenance modelling, offers benefits in terms of assessing compliance with the FAIR provenance definition, as well as providing transparent information for better analysis.

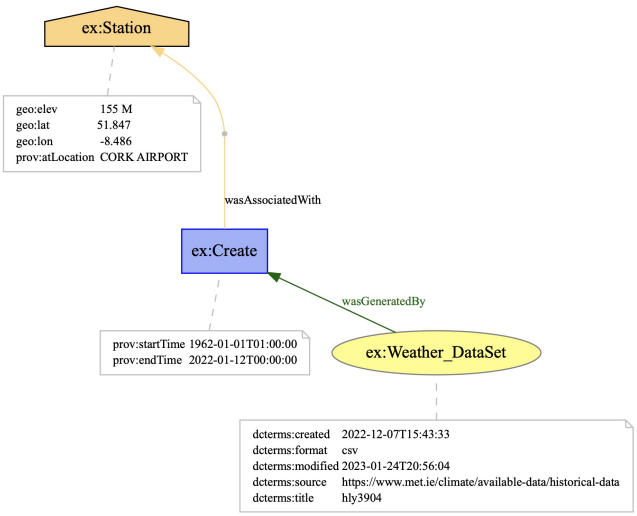

Figure 4: Provenance of Cork airport station.

Manually collecting provenance records or searching for them in open repositories and/or datasets can be time-consuming, making both reusability and reproducibility challenging, as provenance is an essential requirement. Our proposed approach addresses this challenge by providing tailored traceability and visualization. Our

output provenance graph offers information that supports the reuse and reproduction of weather time series datasets, directly visualized in the graph. For example, a scientist interested in analysing a weather station's records, with particular specifications regarding station height, longitude, latitude, and location, could use the graph to make decisions about the intended dataset. Similarly, provenance properties such as the interval time of the dataset, its source, and its file format can be analysed depending on the research question. This comprehensive view of the input file containing the time series enhances transparency and reproducibility.

Scientists can track previously downloaded time series using the graphical information provided in Figure 5. The figure depicts a use case where a scientist wishes to perform an exploratory study on the air temperature of the northwest of Ireland. The first stage involves determining which datasets should be used, based on location and other criteria. The graph templates provide geospatial information such as longitude, altitude, and height. Historical time series data are often required for forecasting studies, and the selected dates and times need to be identified. Depending on the research, one can analyse datasets from the same resource or from different ones.

It is also noted that the provenance graphs can help to understand the complete lifecycle of the weather time series. For each dataset, we can check the resource link from where it was downloaded as shown in Figure 3 and Figure 4, where both have resources. That is to say, provenance graphs provide a complete lifecycle of the weather time series, enabling researchers to understand when and where the data was collected, and track when a time series was created and updated. This information is useful for reusability and reproducibility, as it enables researchers to assess the origin and history of the data, and ensure that it is appropriate for their research.

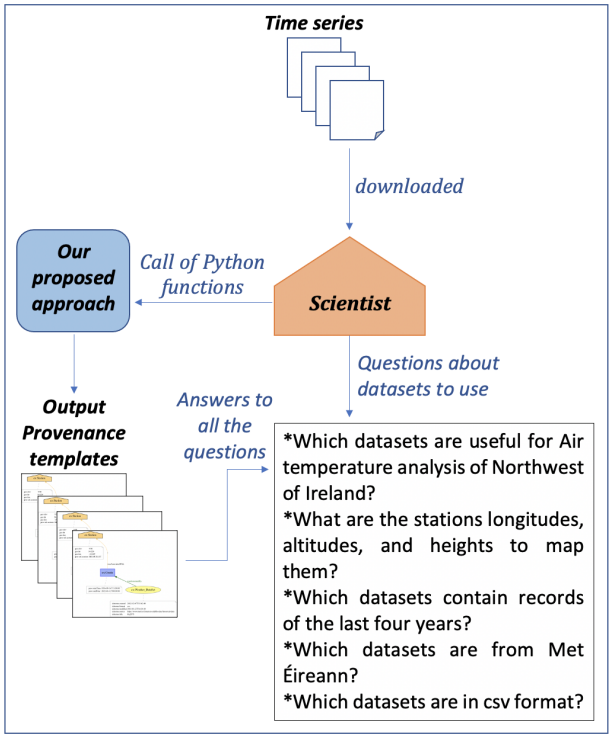

Figure 5: Effectiveness of Provenance templates outputs

5 CONCLUSIONS

One of the biggest challenges in promoting dataset reusability and reproducibility is the lack of complete metadata information. In this paper, we showcase the effectiveness of our approach in a real-world use case using open weather station records. Our approach utilises graph templates to clearly assemble the relevant records held by each entity, as demonstrated through the examples of Belmullet and Cork airport.

This example is particularly helpful for climate researchers and scientists who wish to compare weather station records without having to repeatedly access resources and manually retrieve distributed metadata for each time series. More generically, researchers can also analyse time series datasets with similar characteristics–such as location or proximity of stations for our example–which could lead to new research questions, insights, and investigations. Therefore, our approach plays an important role in promoting reusability and reproducibility.

Our approach is based on the PROV-DM model, which allows for the explicit representation of time series and their relationships. We demonstrate the automatic creation of PROV-DM graphs from publicly accessible resources and illustrate our approach using two different weather time series. We believe this research perspective could be expanded to assess the provenance compliance of other time series resources with FAIR principles. Collecting essential metadata properties would enable FAIR provenance scoring. Our present paper encourages further research in scoring FAIR principles.

Our future work will focus on metadata provenance subsets and evaluating the provenance model on other dataset types, resources, and domains. We aim to investigate scenarios related to datasets of different versions within various timestamps using the weather meta-dataset provenance. Additionally, we plan to carry out more use case studies across various research domains and collect relevant provenance metadata using our proposed Provenance data model graph, which can be extended or readjusted to comply with FAIR guidelines. Finally, we plan to integrate Python functions as a service with other repositories.

ACKNOWLEDGMENTS

This work has been partly developed under the auspice of the projects *i*) "TrainRDM Open Science and Research Data Management Innovative and Distributed Training Programme" funded from Nov/2020 to Apr/2023 by the European Commission Erasmus+ Programme under Grant No.: 2020-1-RO01-KA203-080170; and, *ii*) "Smardy: Marketplace for Technology Transfer of Research Data, Software, and Results" (2021–2024) funded by the European Eureka Network through Ireland's International Research Fund of Enterprise Ireland (Grant No.: IR20210058).

REFERENCES

[1] C. A. Anderson. Temperature and aggression: Ubiquitous effects of heat on occurrence of human violence. *Psychological Bulletin*, 106(1):74–96, 1989.

[2] R. Ashrafi, M. Amirahmadi, M. Tolou-Askari, and V. Ghods. Multi-objective resilience enhancement program in smart grids during extreme weather conditions. *International Journal of Electrical Power and Energy Systems*, 129:106824, July 2021.

[3] M. Baker. 1, 500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, May 2016.

[4] B. V. Batlajery. Modeling and generating marketplace activities with PROV-DM. In *AI4I*, pages 91–94, Laguna Hills, Sept. 2021. IEEE.

[5] A. Chebotko, J. Abraham, P. Brazier, A. Piazza, A. Kashlev, and S. Lu. Storing, indexing and querying large provenance data sets as RDF graphs in apache HBase. In *2013 IEEE Ninth World Congress on Services*, pages 1–8, Santa Clara, June 2013. IEEE.

[6] E. D. Clarke, S. Griffin, F. McDermott, J. M. Correia, and C. Sweeney. Which reanalysis dataset should we use for renewable energy analysis in ireland? *Atmosphere*, 12(5):624, May 2021.

[7] D. L. da Silva, A. Batista, and P. L. P. Correa. Data provenance in environmental monitoring. In *MASS*, pages 337–342, Brasilia, Oct. 2016. IEEE.

[8] S. B. Davidson and J. Freire. Provenance and scientific workflows: Challenges and opportunities. In *SIGMOD '08*, pages 1345–1350, Vancouver, 2008. Association for Computing Machinery.

[9] J. J. A. Denissen, L. Butalid, L. Penke, and M. A. G. van Aken. The effects of weather on daily mood: A multilevel approach. *Emotion*, 8(5):662–667, 2008.

[10] S. S. Edit Herczog, Keith Russell. FAIR data maturity model: specification and guidelines. Technical report, Research Data Alliance, https://doi.org/10.15497/RDA00050, June 2020.

[11] I.-D. Filip, C. Ionite, A. González-Cebrián, M. Balanescu, C. Dobre, A. E. Chis, D. Feenan, A.-A. Buga, I.-M. Constantin, G. Suciu, G. V. Iordache, and H. González-Vélez. SMARDY: Zero-trust FAIR marketplace for research data. In *IEEE BigData*, pages 1535–1541, Osaka, Dec. 2022. IEEE.

[12] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers. Examining the challenges of scientific workflows. *Computer*, 40(12):24–32, Dec. 2007.

[13] H. González-Vélez, C. Dobre, B. S. Solís, G. Antinucci, D. Feenan, and D. Gheorghe. Open science and research data management: A FAIR European postgraduate programme. In *IEEE BigData*, pages 2522–2531, Osaka, Dec. 2022. IEEE.

[14] P. C. Griffin, J. Khadake, K. S. LeMay, S. E. Lewis, et al. Best practice data life cycle approaches for the life sciences. *F1000Research*, 6:1618, June 2018.

[15] P. Groth, H. Cousijn, T. Clark, and C. Goble. FAIR data reuse – the path through data citation. *Data Intelligence*, 2(1-2):78–86, Jan. 2020.

[16] M. Herschel, R. Diestelkämper, and H. B. Lahmar. A survey on provenance: What for? what form? what from? *The VLDB Journal*, 26(6):881–906, Oct. 2017.

[17] D. Hirshleifer and T. Shumway. Good day sunshine: Stock returns and the weather. *The Journal of Finance*, 58(3):1009–1032, May 2003.

[18] A. Jacobsen, R. Kaliyaperumal, L. O. B. da Silva Santos, B. Mons, E. Schultes, M. Roos, and M. Thompson. A generic workflow for the data FAIRification process. *Data Intelligence*, 2(1-2):56–65, Jan. 2020.

[19] A. Kyrtsoglou, D. Asimina, D. Triantafyllidis, S. Krinidis, K. Kitsikoudis, D. Ioannidis, S. Antypas, G. Tsoukos, and D. Tzovaras. Missing data imputation and meta-analysis on correlation of spatio-temporal weather series data. In *IEMCON*, pages 496–502, Vancouver, Oct. 2021. IEEE.

[20] R. P. Larrick, T. A. Timmerman, A. M. Carton, and J. Abrevaya. Temper, temperature, and temptation. *Psychological Science*, 22(4):423–428, Feb. 2011.

[21] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L. O. B. da Silva Santos, and M. D. Wilkinson. Cloudy, increasingly FAIR: revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services and Use*, 37(1):49–56, Mar. 2017.

[22] L. Moreau, B. V. Batlajery, T. D. Huynh, D. Michaelides, and H. Packer. A templating system to generate provenance. *IEEE Transactions on Software Engineering*, 44(2):103–121, Feb. 2018.

[23] L. Moreau, P. Groth, J. Cheney, T. Lebo, and S. Miles. The rationale of PROV. *Journal of Web Semantics*, 35:235–257, 2015.

[24] L. Moreau and P. Missier. PROV-DM: The PROV Data Model. W3C Recommendation 30 April 2013, World Wide Web Consortium (W3C), http://www.w3.org/TR/prov-dm/, 2013. (Last accessed: 28/Jan/2023).

[25] E. Nwafor, A. Campbell, D. Hill, and G. Bloom. Towards a provenance collection framework for internet of things devices. In *SmartWorld*, pages 1–6, San Francisco, Aug. 2017. IEEE.

[26] M. Sampaio, A. L. Ferreira, J. A. Castro, and C. Ribeiro. Training biomedical researchers in metadata with a MIBBI-based ontology. In *MTSR*, volume 1057 of *Communications in Computer and Information Science*, pages 28–39. Springer, Rome, 2019.

[27] M. Stoffers, M. Meinel, B. Hofmann, and A. Schreiber. Integrating provenance-awareness into the space debris processing system BACARDI. In *2022 AERO*, pages 1–12, Big Sky, Mar. 2022. IEEE.

[28] J. Top, S. Janssen, H. Boogaard, R. Knapen, and G. Şimşek-Şenel. Cultivating FAIR principles for agri-food data. *Computers and Electronics in Agriculture*, 196:106909, May 2022.

[29] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018:1–9, Mar. 2016.

[30] M. D. Wilkinson, M. Dumontier, S.-A. Sansone, et al. Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Scientific Data*, 6(1):174:1–12, Sept. 2019.