

A Systematic Literature Review on Numerical Weather Prediction Models and Provenance Data

Alper Tufek(✉) and Mehmet S. Aktas

Computer Engineering Department, Yildiz Technical University, Istanbul, Turkey
alper.tufek@std.yildiz.edu.tr, aktas@yildiz.edu.tr

Abstract. The weather has been an important issue for mankind since the earliest times. The need to predict the weather accurately increases every day when considering the effects of natural disasters such as floods, hails, extreme winds, landslides, etc. on many sectors from transportation to agriculture, which all depend on weather conditions. Numerical weather prediction (NWP) models, today's the de-facto tools used for weather forecasting, are scientific software that models atmospheric dynamics in accordance with the laws of physics. These models perform complex mathematical calculations on very large data (gridded) and require high computational power. For this reason, NWP models are scientific software that is usually run on distributed infrastructures and often takes hours to finish. On the other hand, provenance is another key concept as important as weather prediction. Provenance can be briefly defined as metadata that provides information about any kind of data, process, or workflow. In this SLR study, a comprehensive screening of literature was performed to discover primary studies that directly suggest systematic provenance structures for NWP models, or primary studies in which at least a case study was implemented on an NWP model even if considered in a broader perspective. Afterward, these primary studies were thoroughly examined in line with specific research questions, and the findings were presented in a systematic manner. An SLR study on primary studies which combines the two domains of NWP models and provenance research has never been done before. So we think that this work will fill an important gap in literature regarding studies combining the two domains and increase the interest in the subject.

Keywords: Numerical weather prediction (NWP) · Systematic Literature Review (SLR) · Provenance

1 Introduction

The weather has been an important phenomenon for mankind since the earliest times. It is almost impossible to find an area where the weather conditions are not effective. The magnitude of the need to predict the weather will be

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13380, pp. 616–627, 2022.
https://doi.org/10.1007/978-3-031-10542-5_42

better understood when considering the effects of natural disasters such as floods, hails, extreme winds, landslides, etc. on many sectors from transportation to agriculture, which all depend on weather conditions. It is for this reason that since ancient times, people have tried to predict the weather by observing various patterns in clouds or different natural events.

Today, the de-facto tools used for weather forecasting are numerical weather prediction (NWP) models. These models are scientific software that models atmospheric dynamics in accordance with the laws of physics. Their working logic is based on simulating the movement of air according to the laws of fluid dynamics and physics, and calculating the possible states of the atmosphere in future time periods, taking the current (actual) state of the atmosphere as the starting point. This is a process that performs complex mathematical calculations on very large data (gridded) and therefore requires high computational power. For this reason, NWP models are scientific software that is usually run on distributed infrastructures and often takes a long time to terminate.

On the other hand, provenance is another key concept as important as weather prediction. Provenance can be briefly defined as metadata that provides information about any kind of data, process, or workflow. For example, if it is a data product, data provenance can provide very detailed information about what processes are employed with which parameters during the production of that data from its source. Or, in situations where various simulations are run through scientific workflow applications, runtime provenance can provide useful information about all workflow stages (call traces, input parameters, artifacts, etc.).

Although there are survey or review studies on provenance systems in the literature [2,12,15] we did not find any SLR or mapping study specific to numerical weather forecasting models from the perspective of provenance. In this SLR study, a comprehensive screening study was aimed at primary studies that directly suggest systematic provenance structures for NWP models, or at the primary studies in which at least a case study was implemented on an NWP model even if considered in a broader perspective.

2 Research Methodology

In this SLR study, we followed the guidelines proposed in [13] and [6]. In this direction, primary studies on the subject were scanned in the literature. A primary study is an experimental-based study that investigates a specific research question in any field, whereas a secondary study is one in which all primary studies related to a specific research question are reviewed and the findings are compiled and synthesized.

2.1 Research Questions

In this study, we focus on primary studies that directly suggest systematic provenance structures for NWP models, or primary studies in which at least a case study was implemented on an NWP model even if considered in a broader perspective. We asked various questions to direct the knowledge synthesis. Our research questions and the main motivations behind them are shown in Table 1.

Table 1. Research questions

Research question	Main motivation
RQ1: What provenance standards/specifications are proposed/used?	PROV, OPM, or a specific representation/format
RQ2: What are the provenance sources proposed?	From where provenance is obtained: Log files, workflow app, etc.?
RQ3: What provenance storage mechanisms are proposed?	Special provenance storage system, file, or database, etc.
RQ4: What is the distribution of publications by years?	To show the trend analysis
RQ5: In which journals or conferences have related papers been published?	To identify journals and conferences where researchers publish their studies on the topic

2.2 Search Strategy

In this SLR, we specifically focus on provenance data within NWP models. We performed specific searches on all best known digital libraries to gather the primary studies [13]. These libraries are ACM Digital Library, IEEE Xplore, Scopus, SpringerLink, Elsevier's Science Direct and Web of Science. Each library has its own search engine with its own characteristics. We tried to be as much inclusive as possible when constructing query string. Numerical weather prediction models may be referred to as "atmospheric modeling", "numerical weather forecasting", etc. If supported by the search engine of the digital source, wildcards are used to include every different form of keywords, otherwise most possible variations of keywords are exclusively included in query string. The query strings used for each library and quick notes about each search engine are provided in Table 2.

Since the first Working Draft of the PROV Family of Documents [14] was published in December 2012 by W3C, papers with a publication date prior to 2012 were discarded. Also, the inclusion and exclusion criteria were defined to identify relevant works focused on in this SLR.

Inclusion Criteria (IC):

- **IC1:** Research articles or short papers that focus on provenance data in NWP workflows.

Exclusion Criteria (EC):

- **EC1:** Duplicated papers;
- **EC2:** Papers not in English;
- **EC3:** Surveys, Review Articles, Book Chapters, Posters, Tutorials, Technical Reports and Standards;

During the SLR study, StArt [16] and JabRef [5] tools were used. StArt is a free tool that aids performing SLR studies in accordance with the systematic review procedure suggested by Kitchenham [6]. JabRef (another free tool) is a reference management software that can import more than 15 reference formats, and it supports many operations such as merge, convert, export, etc. Search results obtained from digital libraries are first exported to files in bibtex, RIS or CSV formats depending on the support provided by the search engine. Next,

Table 2. Data sources and query strings used

Source	Query string	Notes
ACM	(NWP OR (numerical AND (weather OR atmospher*) AND (prediction OR forecast* OR model*))) AND provenance	Supports wildcards and boolean operators such as AND, OR, and NOT
IEEE	(NWP OR (numerical AND (weather OR atmospher*) AND (prediction OR forecast* OR model*))) AND provenance	Wildcards and boolean operators are supported. Search in Full Text & Metadata was selected
Scopus	(NWP OR (numerical AND (weather OR atmospher*) AND (prediction OR forecast* OR model*))) AND provenance	Wildcards and boolean connectors are supported
SpringerLink	(NWP OR (numerical AND (weather OR atmospher*) AND (prediction OR forecast* OR model*))) AND provenance	Wildcards and boolean connectors are supported
Science Direct	(NWP OR (numerical (weather OR atmosphere OR atmospheric) (prediction OR forecast OR forecasting OR model OR modeling))) AND provenance	Wildcards are not supported. Maximum of 8 boolean connectors are allowed. Spaces between keywords or keyword groups are implicitly interpreted as AND connector. Plurals and spelling variants are included (colour vs color, attack vs attacks)
Web of Science	(NWP OR (numerical AND (weather OR atmospher*) AND (prediction OR forecast* OR model*))) AND provenance	It supports wildcards and boolean operators. Stemming and lemmatization are applied by default to expand the search

these files were imported into the JabRef tool, grouped as a single file for each data source. Then each file per data source is imported into the corresponding search session defined for each data source in the StArt tool. With the StArt tool, it is possible to list all primary sources together, and to easily classify them as Accepted or Rejected by quickly evaluating titles, keywords, and abstracts if available.

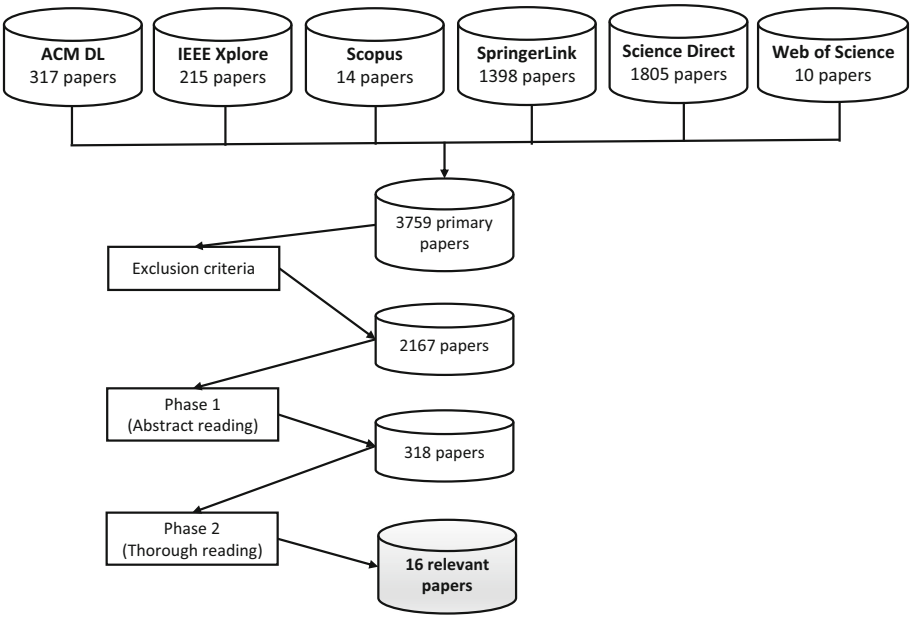

Fig. 1. Paper selection flow and summary statistics

3 Results and Analysis

We performed the extraction in January 2022. A total of 3759 results were obtained when a search was made on all search engines with the constraint of the publication years between 2012 and 2021 inclusive. After applying the exclusion criteria, there were 2176 results left (duplicate results, surveys, review articles, book chapters, posters, tutorials, etc. were removed). Then, in the next step (Phase 1), titles, keywords and abstracts were read and irrelevant articles were eliminated. At the end of this phase, there were 318 studies left. At the last stage, all the remaining articles were examined in depth and relevant papers within the scope of this SLR were identified and the findings obtained in line with the research questions were synthesized. The workflow for paper selection and the number of primary studies obtained at each stage are summarized in Fig. 1.

The findings obtained from the relevant studies and the evaluation of the synthesis results for each research question are as follows.

3.1 RQ1: What Provenance Standards/specifications Are Proposed/Used?

In most of the selected studies, OPM [9] and its improved version, PROV specification [8], were used. However, apart from these, some studies use a provenance format of their own, while some other studies use various known but less commonly used provenance formats. The provenance standards used/supported in the selected primary studies, and their explanations are summarized in the Table 3.

Table 3. Provenance standards/specifications that are proposed/used in primary studies

Study ID	Provenance standards/specifications used
S01	EML - Ecological Metadata Language
S02, S03, S04	Open Provenance Model (OPM) [9]
S05, S06, S07, S08, S14	No known provenance format specified
S09, S10, S11, S12	PROV
S13	OWL and RDF
S15	Provenance-Aware Storage System, PASS [10,11] and OPM
S16	Two extension schemas of PROV-O; CWLProv (for file-oriented provenance) and S-Prov (for streaming data provenance)

3.2 RQ2: What Are the Provenance Sources Proposed?

The provenance sources used/proposed in the selected primary studies are summarized in the Table 4. It is possible to use a variety of sources to obtain structured provenance. Scientific software is often run in a certain order according to a defined workflow through a workflow manager. Therefore, the most common source of provenance is the tracks that workflow management software keeps while running tasks within the workflow. However, in this method, there may be provenance losses for the intermediate processes of an atomic program in the workflow. Another important source of provenance is the log files created directly by the scientific software itself. In this method, it is possible to capture the details of intermediate transactions, but on the other hand, the fact that there is not a common standard for log file formats brings additional difficulties. Yet another method is the instrumentation of the source codes of the software so that the application itself generates structured provenance directly at runtime.

Each method has its own advantages and disadvantages. For example, in the source code instrumentation method, no intermediate extractor will be needed, but it is not always possible to find the source code of the software. Even if the source code is available, making additional changes to the code may cause the program to exhibit ambiguous behavior or introduce bugs.

Table 4. Provenance sources that are proposed/used in primary studies

Study ID	Provenance sources used
S01, S05, S06, S08, S13, S14	Workflow runtime provenance
S02, S03, S04	Provenance dataset consisting of provenance from different types of workflows (approximately 48,000 workflow execution instances) [9]
S07	Their private steering engine called Co^2Steer
S09	Their framework named CityPulse
S10, S11, S12	Log files
S15	Workflow traces from PASS, Karma, Tupelo [3], Taverna [7], PASOA [4]
S16	Provenance traces produced by scientific workflows of two major types: file-oriented and data-streaming. They are obtained from SPARQL endpoints [19]

3.3 RQ3: What Provenance Storage Mechanisms are Proposed?

The provenance storage mechanisms used/proposed in the selected primary studies, and their explanations are summarized in the Table 5. The most well-known system developed specifically for structured provenance storage is Karma [17]. While Karma supports the OPM standard, Komadu [18], another provenance system, which is an advanced version of Karma, fully supports PROV specification. Apart from these two systems, there are also studies that directly use a relational database (MySQL, BerkeleyDB, etc.). While some studies do not give details about the provenance storage environment, in one study, the Kepler [1] workflow manager was extended and provenance support was added.

3.4 RQ4: What is the Distribution of Publications by Years?

Looking at the final extraction results, we can say that there is not much relevant studies on the topic. This fact can be clearly seen in Fig. 2. Although there are many primary studies in the NWP domain, there are not many studies combining the NWP domain and the concept of systematic provenance, which is mostly

Table 5. Provenance storage mechanisms that are used in primary studies

Study ID	Provenance storage mechanism(s) used
S01	A provenance storage system which extends the Kepler [1] scientific workflow system
S02, S03, S04	Karma [17] provenance system
S05	A provenance module named ProvEn developed specifically for The Earth System Grid Federation (ESGF)
S06, S07, S08, S09, S14, S16	Uses their own internal databases whose details are not given in the paper
S10, S11, S12	Komadu [18] (advanced version of the Karma)
S13	Relational database (i.e. MySQL) or XML files
S15	A BerkeleyDB database storing key/value data

researched in the field of software engineering. On the other hand, although there is no general pattern in the distribution across years, we can say that there has been a relative increase in the number of studies on the subject as of 2021. With this SLR study, we think that we will fill an important gap in the literature regarding studies combining the two domains and increase the interest in the subject.

3.5 RQ5: In Which Journals or Conferences Have Related Papers Been Published?

The list of journals or conferences in which relevant studies are published can be shown in Table 6.

Since the total number of relevant studies was small, we did not expect a clustering in on single journal or conference. Indeed, only one paper per journal or conference has been published in across all journals/conferences except for the Future Generation Computer Systems journal. Three different studies have been published in the Future Generation Computer Systems journal.

4 Threats to Validity of Research

The possible threats that may affect the validity of the results obtained in this SLR study and the precautions we tried to take in response to these threats are as follows:

Selection Bias: This threat refers to the incorrect inclusion/exclusion of some primary studies in/from the SLR. In order to eliminate this possibility as much as possible, a very rigorous selection procedure has been followed and the StArt tool has been used for this end. The StArt tool facilitates the evaluation of primary studies obtained from all data sources by importing all bibliographic records into a single place. In addition, its ability to detect duplicate studies and

Table 6. Journals/conferences where related papers published

Journal/Conference name	Type	Notes
ACM Transactions on Storage (TOS)	Journal	ACM journal
Concurrency and Computation: Practice and Experience	Journal	Wiley journal
Environmental Modelling and Software	Journal	Hosted on ScienceDirect
European Conference on Parallel Processing (Euro-Par)	Conference	Covers all aspects of parallel and distributed processing
Future Generation Computer Systems (FGCS)	Journal	Hosted on ScienceDirect
IEEE Access	Journal	Open access (OA)
IEEE International Conference on Big Data and Cloud Computing (BdCloud)	Conference	Annual conference
IEEE Transactions on Cloud Computing	Journal	Dedicated to cloud computing
International Conference on Computational Science (ICCS)	Conference	Annual conference
International Conference on E-Science	Conference	Annual conference
International Conference on Semantics, Knowledge and Grid (SKG)	Conference	Annual conference
Journal of Data and Information Quality	Journal	ACM journal
Proceedings of the ACM on Human-Computer Interaction	Journal	ACM journal
Proceedings of the VLDB Endowment	Journal	ACM journal

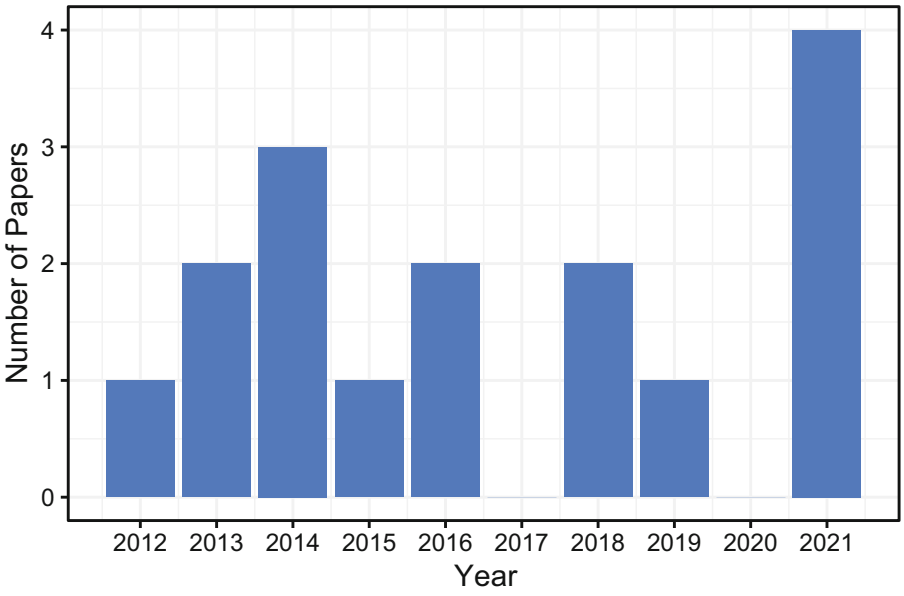

Fig. 2. The distribution of publications across years

the support for performing the selection and extraction phases separately and sequentially reduces the risk of error in the execution of the SLR process.

Publication Selection: We performed searches in all major academic data sources to avoid overlooking any studies that might fall within the scope of this SLR. While performing these searches, we preferred to query all fields (title, keywords, abstract, etc.) and full text whenever supported by the search engine. Besides, we used wildcards to include every different form of keywords if supported, otherwise we exclusively included most possible variations of keywords in query strings.

5 Conclusions

We conducted a systematic literature review to frame the studies that somehow combines the two domains of NWP models and provenance research. We identified 16 relevant papers published between 2012 and 2021. To our knowledge, this kind of an SLR study has never been done before. So we think that this work will fill an important gap in literature regarding studies combining the two domains and increase the interest in the subject.

A Appendix: Relevant Primary Studies

S01	Behrens, H. W., Candan, K. S., Chen, X., Gadkari, A., Garg, Y., Li, M. L., ..., Sapino, M. L. Datastorm-FE: A data-and decision-flow and coordination engine for coupled simulation ensembles. Proceedings of the VLDB Endowment, 11(12), 1906–1909 (2018)
S02	Cheah, Y. W., Plale, B. Provenance quality assessment methodology and framework. Journal of Data and Information Quality (JDIQ), 5(3), 1–20 (2014)
S03	Chen, P., Plale, B., Aktas, M. S. Temporal representation for scientific data provenance. In 2012 IEEE 8th International Conference on E-Science, IEEE, 1–8 (2012)
S04	Chen, P., Plale, B., Aktas, M. S. Temporal representation for mining scientific data provenance. Future generation computer systems, 36, 363–378 (2014)
S05	Cinquini, L., Crichton, D., Mattmann, C., Harney, J., Shipman, G., Wang, F., ..., Schweitzer, R. The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data. Future Generation Computer Systems, 36, 400–417 (2014)
S06	Galizia, A., Roverelli, L., Zereik, G., Danovaro, E., Clematis, A., D'Agostino, D. Using Apache Airavata and EasyGateway for the creation of complex science gateway front-end. Future Generation Computer Systems, 94, 910–919 (2019)
S07	Liu, W., Ye, Q., Wu, C. Q., Liu, Y., Zhou, X., Shan, Y. Machine Learning-assisted Computational Steering of Large-scale Scientific Simulations. In 2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), 984–992 (2021)
S08	Lopez, J. L. A., Kalyuzhnaya, A. V., Kosukhin, S. S., Ivanov, S. V. Data quality control for St. Petersburg flood warning system. Procedia Computer Science, 80, 2128–2140 (2016)
S09	Puiu, D., Barnaghi, P., Tönjes, R., Kümper, D., Ali, M. I., Mileo, A., ..., Fernandes, J. Citypulse: Large scale data analytics framework for smart cities. IEEE Access, 4, 1086–1108 (2016)
S10	Tufek, A., Gurbuz, A., Ekuklu, O. F., Aktas, M. S. Provenance collection platform for the weather research and forecasting model. In 2018 14th International Conference on Semantics, Knowledge and Grids (SKG), IEEE, 17–24 (2018)
S11	Tufek, A., Aktas, M. S. On the provenance extraction techniques from large scale log files: a case study for the numerical weather prediction models. In European Conference on Parallel Processing, Springer, 249–260 (2020)
S12	Tufek, A, Aktas, MS. On the provenance extraction techniques from large scale log files. Concurrency and Computation: Practice and Experience (2021). https://doi.org/10.1002/cpe.6559
S13	Turuncoglu, U. U., Dalfes, N., Murphy, S., DeLuca, C. Toward self-describing and workflow integrated Earth system models: A coupled atmosphere-ocean modeling system application. Environmental modelling & software, 39, 247–262 (2013)
S14	Wu, C. Q., Lin, X., Yu, D., Xu, W., Li, L. End-to-end delay minimization for scientific workflows in clouds under budget constraint. IEEE Transactions on Cloud Computing, 3(2), 169–181 (2014)
S15	Xie, Y., Muniswamy-Reddy, K. K., Feng, D., Li, Y., Long, D. D. Evaluation of a hybrid approach for efficient provenance storage. ACM Transactions on Storage (TOS), 9(4), 1–29 (2013)
S16	Zhao, R., Atkinson, M., Papapanagiotou, P., Magnoni, F., Fleuriot, J. Dr. Aid: Supporting Data-governance Rule Compliance for Decentralized Collaboration in an Automated Way. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1–43 (2021)

References

1. Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B., Mock, S.: Kepler: an extensible system for design and execution of scientific workflows. In: Proceedings. 16th International Conference on Scientific and Statistical Database Management, pp. 423–424. IEEE (2004)
2. Freire, J., Koop, D., Santos, E., Silva, C.T.: Provenance for computational tasks: a survey. Comput. Sci. Eng. **10**(3), 11–21 (2008)
3. Futrelle, J., et al.: Semantic middleware for e-science knowledge spaces. In: Proceedings of the 7th International Workshop on Middleware for Grids, Clouds and e-Science (2009)
4. Groth, P., et al.: An architecture for provenance system. Technical report (2006)
5. JabRef - Free JabRef Reference Management Tool. https://www.jabref.org. Accessed January 2022
6. Kitchenham, B., Charters, S.: Guidelines for performing systematic literature reviews in software engineering (2007)
7. Missier, P., et al.: Taverna, reloaded. In: Proceedings of the 22nd International Conference on Scientific and Statistical Database Nanagement (SSDBM 2010) (2010)
8. Missier, P., Belhajjame, K., Cheney, J.: The W3C PROV family of specifications for modeling provenance metadata. In: Proceedings of the 16th International Conference on Extending Database Technology, pp. 773–776 (2013). https://doi.org/10.1145/2452376.2452478
9. Moreau, L., et al.: The open provenance model core specification (v1.1). Futur. Gener. Comput. Syst. **27**(6), 743–756 (2011)
10. Muniswamy-Reddy, K.K., Holland, D.A., Braun, U., Seltzer, M.I.: Provenance-aware storage systems. In: USENIX Annual Technical Conference, General Track, pp. 43–56 (2006)
11. Muniswamy-Reddy, K.K., et al.: Layering in provenance systems. In: Proceedings of the 2009 USENIX Annual Technical Conference (USENIX 2009). USENIX Association (2009)
12. Pérez, B., Rubio, J., Sáenz-Adán, C.: A systematic review of provenance systems. Knowl. Inf. Syst. **57**(3), 495–543 (2018)
13. Petersen, K., Feldt, R., Mujtaba, S., Mattsson, M.: Systematic mapping studies in software engineering. In: Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering (EASE 2008) (2008)
14. PROV-Overview. https://www.w3.org/TR/2012/WD-prov-overview-20121211/. Accessed 20 Dec 2021
15. Simmhan, Y.L., Plale, B., Gannon, D.: A survey of data provenance in e-science. ACM SIGMOD Rec. **34**(3), 31–36 (2005)
16. StArt - State of the Art through Systematic Review. LAPES Laboratory of Research on Software Engineering (LAPES). http://lapes.dc.ufscar.br/tools/start_tool. Accessed 10 Jan 2022
17. Simmhan, Y.L., Plale, B., Gannon, D.: Karma2: provenance management for datadriven workflows. Int. J. Web Serv. Res. (IJWSR) **5**(2), 1–22 (2008)
18. Suriarachchi, I., Zhou, Q., Plale, B.: Komadu: a capture and visualization system for scientific data provenance. J. Open Res. Softw. **3**(1) (2015)
19. W3C: SPARQL 1.1 Query Language (2013). https://www.w3.org/TR/sparql11-query/