2022 IEEE/ACM Workshop on Workflows in Support of Large-Scale Science (WORKS) | 978-1-6654-5191-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/WORKS56498.2022.00007

Challenges of Provenance in Scientific Workflow Management Systems

Khairul Alam
Department of Computer Science
University of Saskatchewan
Saskatoon, Canada
kha060@usask.ca

Banani Roy
Department of Computer Science
University of Saskatchewan
Saskatoon, Canada
banani.roy@usask.ca

***Abstract*—Scientific workflow is one of the well-established pillars of large-scale computational science and emerged as a torchbearer to formalize and structure a massive amount of complex heterogeneous data and accelerate scientific progress. A workflow can analyze terabyte-scale datasets, contain numerous individual tasks, and coordinate between heterogeneous tasks with the help of scientific workflow management systems (SWfMSs). SWfMSs support the automation of repetitive tasks and capture complex analysis through workflows. However, the execution of workflows is costly and requires a lot of resource usage. At different phases of a workflow life cycle, most SWfMSs store provenance information, allowing result reproducibility, sharing, and knowledge reuse in the scientific community. But, this provenance information can be many times larger than the workflow and input data, and managing provenance data is growing in complexity with large-scale applications. Handling exponential increasing data volume and utilizing the technical resources for storage and computing are thus demanded by exploiting data-intensive computing in various application fields. This paper documented the challenges of provenance management and reuse in e-science, focusing primarily on scientific workflow approaches by exploring different SWfMSs and provenance management systems. We also investigated the ways to overcome the challenges.**

***Index Terms*—Scientific workflow, scientific workflow management system, provenance, reusability, open science**

I. Introduction

Scientific workflows (shortened as workflows in this work) and scientific workflow management systems (SWfMSs) provide potential opportunities to solve complex, multidisciplinary, and large-scale computational experiments like Cybershake [32], Montage [61], Epigenomics [80], coronavirus sequencing [81], cancer studies [83], molecular dynamics [84], and so on. However, many of these workflows have significant computational, storage, and communication demands and thus must execute on a wide range of large-scale platforms [1], from large clouds to upcoming exascale HPC platforms. In SWfMSs, scientists often need to rerun workflows for finer-grained analyses and improve the quality of scientific analyses, leading to the FAIR [74] (*Findable, Accessible, Interoperable and Reusable*) datasets and analysis pipelines. In such frequently executed workflows, processing large datasets that include computationally expensive modules requires a long execution time. The more data to process, the longer the workflow may take, which may be days depending on the problem and HPC environment [60]. Due to the popularity of workflows, hundreds of workflow management systems have been developed, but proper management (storing and reusing) of provenance information is still not up to the mark. Furthermore, given the increasing number of high-quality public datasets and pipelines, this lack of clear compatibility threatens the findability and reusability of these resources [8].

SWfMSs are becoming popular due to their capacity to manage complex and diverse applications, and this has paved the way for scientists to accelerate many scientific discoveries. However, if the same workflow or sub-workflow needs to be executed several times, it will generate a high volume of undesired redundant provenance data. Due to the complex heterogeneous nature of data and large volume, a single experiment may demand a week to run, even in high-performance computing environments [6]. Handling such exponentially increasing data volumes and utilizing the technical resources for storage and computing are thus demanded in exploiting data-intensive computing in various application fields. Gathering provenance data from distributed workflows raises new and different challenges [50]. Also, a lack of proper coordination and broadly usable standard components lead to the ad hoc and isolated solution rather than adopting and extending existing solutions, which causes wastage of time and resources [9]. Many existing scientific workflow management systems capture detailed provenance information. Each SWfMS has its particular approach for executing a workflow and capturing its provenance information. This provenance information generally consists of data and process dependencies introduced during a workflow run. It is crucial for enabling scientists to more easily understand, reproduce, and verify scientific results [4]. Detail provenance information allows scientists to audit trails and verify and reproduce their results.

Research [8, 17, 18, 19, 29] has been done to reuse existing workflows in different ways. Also, there is a lot of research [24, 26, 30, 34, 53, 55, 56, 69, 95, 96, 99] about storing, querying provenance information, audit trails using provenance also reproducibility issues of a scientific workflow using provenance data. Workflow provenance assures the reliability and integrity of workflows and the data as they are routed in complex workflows. A proper provenance record is essential in many scientific experiments as it enables experiments

978-1-6654-5191-8/22/$31.00 ©2022 IEEE
DOI 10.1109/WORKS56498.2022.00007

to be systematically repeated and validated by others. The amount and cost of provenance information can be inversely proportional to the granularity; they can grow to be larger than the data it describes. Although many SWfMS systems have been developed, there is still a marked lack of research investigating (1) *the reusability of provenance information in a scientific workflow execution* and (2) *optimized storing (non-redundant) of detailed provenance information*. In current SWfMSs, if the same workflow/sub-workflow needs to be executed multiple times, the workflow cannot automatically use the previously stored results. It costs a massive amount of storage and execution time and makes the system inefficient to use.

Open science has emerged as a framework for improving the quality of scientific analysis. Transparent, accessible knowledge sharing and collaborative networks are essential components of open science. Scientific workflow communities are also tending toward open science and, as a result, made many workflows and datasets available to the community in different repositories [85, 86, 87, 88, 89, 91, 92, 93, 94]. In addition, some of the repositories [92, 93, 94] also share some provenance information. The future of scientific advancements mostly depends on the ability of scientists to comprehend the vast amount of data currently being produced and acquired. Unfortunately, while public datasets and pipelines proliferate, researchers remain unassisted in creating relevant analyses from these resources that remain largely underutilized [8]. Although a vast amount of information is available to the community, due to several challenges, they are mostly nonreusable, especially for workflow re-execution purposes.

In this paper, we investigated the challenges of provenance data and recorded the actions to overcome them. In particular, we presented techniques for reducing provenance overload and making provenance data more reusable and fine-grained. Furthermore, we identified that if provenance data can be appropriately managed, it will help tremendously to mitigate resources and time usage, eventually reducing cost and making the data analysis process more effective and efficient. Finally, we tried to answer three research questions and, in this way, made three contributions to this paper as follows:

RQ1: What are the provenance challenges in SWfMS?
We checked the provenance capturing mechanisms of several SWfMSs and investigated the challenges, especially regarding reusability and storing approaches.

RQ2: How can we overcome the provenance challenges?
We identified several challenges and recommended several actions to overcome them. We suggested optimized storing, avoiding redundancies, enabling data sharing, and so on.

RQ3: Can we use provenance data in real-time?
While executing a workflow, we proposed to use existing provenance data instead of executing the workflow module wherever possible. It will save workflow execution time, especially for long-running queries, and ensure reusability.

II. Scientific Workflow Management

This section briefly introduces Scientific Workflow, Scientific Workflow Management Systems (SWfMSs), Scientific Workflow Life Cycle, and Scientific Workflow Examples.

A. Scientific Workflow

Scientific workflows assist scientists in efficiently modeling and expressing a scientific experiment's entire data processing activities. We can represent scientific workflow as a directed acyclic graph[1], in which nodes represent data processing activities, and edges correspond to data dependencies. Sometimes scientists define it as a sequence of actions sufficient for many applications. A scientific workflow is the computerized facilitation or automation of a scientific process, in whole or part, which usually streamlines a collection of scientific tasks with data channels and dataflow constructs to automate data computation and analysis to enable and accelerate scientific discovery [3]. It composes a collection of interdependent tasks which acquire, generate, transform or analyze complex datasets [23]. It consists of input data, modules, different parameters, and module invocation functions. Scientific workflow is used to model and run scientific experiments by assembling scientific data processing activities, and it may contain one or more sub-workflow. An SWfMS should allow workflows or any subset of their actions to be reused within the same execution.

B. Scientific Workflow Life Cycle

The scientific workflow life cycle consists of the composition, mapping, execution, and provenance phases. Each stage of the workflow life cycle produces explicit provenance metadata. We described provenance in more detail in the **Provenance** section. Here, we adopted a combination of workflow life cycle views [3, 39, 52] as follows:

- **Composition Phase:** In this phase, scientists create an abstraction of a scientific workflow using a Textual or Graphical User Interface (GUI). They can add data activities or control flow structures to the workflow during composition. They can also collect prospective provenance in the composition phase.
- **Deployment phase:** In the deployment phase, scientists construct a concrete workflow using concrete methods.
- **Execution Phase:** Scientists process the input data and produce output in this stage. Some SWfMSs store provenance data for further debugging or experimenting during this phase. Scientists can collect retrospective provenance and monitor activities in this phase.
- **Analysis Phase**: Scientists can analyze their results in this phase. They can visualize or query the results. Visualization and provenance techniques are rarely used together, but they increase scientists understanding of results. In the analysis phase, users can access both retrospective and prospective provenances.

[1] https://cran.r-project.org/web/packages/ggdag/vignettes/intro-to-dags.html

C. Scientific Workflow Management Systems (SWfMSs)

A Workflow Management System (WfMS) defines, creates, and manages the execution of workflows. An SWfMS is a specialized form of a WfMS designed specifically to compose and execute a series of computational or data manipulation steps, or workflow, in a scientific application [76]. Some mostly used SWfMSs are Galaxy [2], Taverna [75], Askalon [64], Pegasus [7], VisTrails [48], Cluster Flow [10], Kepler [22], VizSciFlow [20], iPlant [49], Swift [37], Triana [16], Chiron [82] etc. For supporting scientific workflow analysis, most SWfMSs support workflow provenance. Based on provenance, scientists can perform different sorts of analyses.

D. Scientific Workflow Examples

Many workflow users are reluctant to release their code and data to date; the community has lacked detailed knowledge of a range of scientific workflows [62], but several workflows are currently available. NASA/IPAC [2] created Montage astronomy workflow [61] for the Pegasus SWfMS as an open source toolkit; it can be executed in grid environments and is used to evaluate workflow algorithms. SciEvol [63] is executed in the Chiron SWfMS. In the bioinformatics domain, SciEvol is a workflow for molecular evolution reconstruction. Some data-intensive workflows in bioinformatics are SciPhylomics [40], SciPPGx [41], SciPhy [42]. All these workflows are executed in SciCumulus [38] SWfMS. There are also some others available workflow like CyberShake [11], Broadband [12], LIGO Inspiral Analysis Workflow [15], Coronavirus Sequencing [81], and SIPHT [14]. Scientists can publish their workflows in WorkflowHub[3], which is a collaborative environment. The other repositories are [85, 87, 88, 91] and so on. Scientists share their workflows to these repositories so that other users can reuse them for solving their problems.

III. Provenance

The provenance is sometimes referred to as audit trail, lineage, and pedigree and contains information about the process and datasets used to derive the data product. Provenance provides essential documentation to preserve data, determine its quality and authorship, reproduce as well as interpret and validate the associated scientific results [70]. Some researchers like [54, 58] describe provenance in terms of primary data while others describe it as metadata [65, 73]. Some other authors [59] argue that provenance differs from other forms of metadata as it is based on relationships among objects. Authors in [70] considered provenance data as the (semi-) or automatically and systematically captured and recorded information that helps users or computing systems to determine the derivation history of a data product, starting on its original sources and ending at a given repository.

Provenance storage is used to register provenance information. It can be stored locally or distributed. Centralized provenance storage keeps data in one single repository at one single location. They are easy to use, maintain, manage and control for security purposes, but the single point of failure is the problem. In distributed provenance storage, multiple interrelated repositories are distributed over a computer network. The storage can be RDBMS or filesystem. Distributed provenance storage can be classified as homogeneous or heterogeneous based on the types of the storage system. The life cycle of provenance is shown in Figure 1. We obtained it from [72]

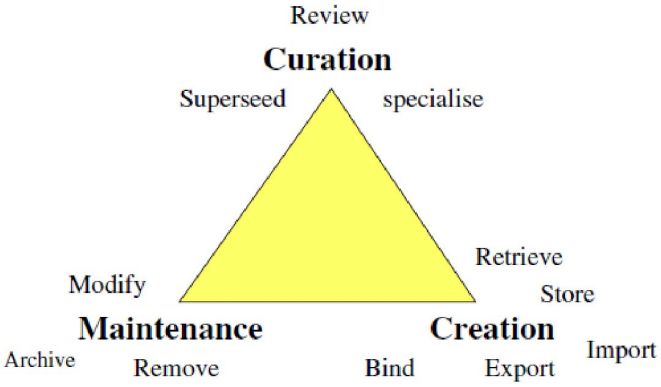

Fig. 1. Provenance Life Cycle

An essential component of provenance is information about causality and another key element is user-defined information which is not captured automatically but records important decisions and notes. Lazy and eager [78] approaches can be used to trace provenance data. Most SWfMSs used eager methods because provenance can be entirely determined here though it adds additional overhead to compute the output and to store annotations in the repository. There are three key components in a provenance management solution: first, the capturing mechanism, then the data model for representing the provenance information, and finally, the infrastructure for storing, accessing, and querying provenance. Two distinct forms of provenance are *prospective* (captures a computational task's specification) and *retrospective* (captures the steps executed as well as information about the environment used to derive a specific data product). In provenance, the dependency relationships among data products and the process that generate them are essential. In SWfMSs, provenance gathering mechanisms are either attached or integrated. They are responsible for storing provenance data using a particular provenance model. Although researchers tried to build a standard open provenance model for capturing the provenance data, provenance is tightly coupled to SWfMSs. Thus, scientific workflows' provenance concepts, representation, and mechanisms differ massively among SWfMSs. SWfMSs capture complex analysis processes at various levels of detail and systematically collect provenance information for the derived data products so they can be queried later.

IV. Related Work

This section discussed related work on scientific workflow management systems provenance based on reusability, storage,

[2]http://montage.ipac.caltech.edu/
[3]https://workflowhub.eu/

and querying systems. At the end of this section, we discussed our research in the context of related work.

A. Storing, querying, and reusing workflow provenance

Our research is a step towards helping scientists diminish their hassle in managing storage and enabling re-usability by skipping executing earlier executed modules. For most scientific workflow management systems, provenance management has become a vital functionality. There are several approaches to managing the provenance of scientific workflows;

RISP [30] technique focused on re-using intermediate stage results generated by previously executed workflows. This technique optimizes storing of workflow data using the mining association rule [28]. A data management scheme to facilitate workflow assembling and executions with different parameter configurations with a GUI is proposed in [33]. [34] focused on easy-to-use and efficient approaches for accessing and querying provenance information. A high-level query language (QLP) is tailored for expressing provenance queries. They provided formal semantics for the language and presented novel techniques, Immediate Lineage Edges (I), Immediate Edges and Dependency Closure (IC), and Immediate Edges and Closure via Pointers (ICP) for efficiently evaluating queries.

Kepler system [68] designed a provenance recording system (*Provenance Recorder (PR)*) to avoid complex configuration process or the scientist implementing a complex API. To give more flexibility, they made their collection facility parametric and customizable. They allowed the user to choose between various levels of detail and even save all of the provenance data needed to recreate a workflow result when the workflow is used as a part of the scientific discovery. Debugging a workflow is also possible in the implementation phase of workflow development. In [69], ProvManager: a provenance management system for scientific workflows is proposed. ProvManager leverages the provenance management at the experiment level by integrating different workflow executions from multiple workflow management systems. It eases the gathering, storing, and analyzing of provenance information in a distributed and heterogeneous environment scenario. As part of the UK's myGrid project, a workflow workbench Taverna [71, 77] is developed. Taverna's provenance model captures locally generated provenance data, and external provenance information gathered from third-party data providers. One advantage of this system is it supports overlaying secondary provenance over the primary logs and lineages. Taverna focused on the semantic web of provenance and showed how it could be mined by a provenance usage component, Provenance Query and Answer (ProQA).

The authors in [24] designed a provenance model that models both prospective and retrospective provenance as an extension to the Open Provenance Model (*OPM*). They implemented a relational provenance store to store, reason, and query prospective and retrospective provenance, which is captured via the proposed provenance collection framework. Provenance trails in the Wings/Pegasus system [55] focused on creating and executing large-scale scientific workflows which involved lots of computations over distributed shared resources. Their approach uses semantic representations to (1) *describe complex scientific applications in a data-independent manner*, (2) *automatically generate workflows of computations for given data sets*, and (3) *map the workflows to available computing resources for efficient execution*. Here, workflow instantiation provenance can be queried using SPARQL [97], and workflow execution provenance can be queried using SQL. The PReServ/PASOA system [53, 56] supports the recording of interaction provenance, actor provenance, and input provenance with the provenance recording protocol. Provenance capturing through operating system using Berkeley DB and XML database is focused on PASS [25] and Earth System Science Server (ES3)[35].

[26] introduced a layered model to represent workflow provenance that allows navigation from an abstract model of the experiment to instance data collected during a specific experiment run. They developed a method called *REDUX* to explore the benefits and challenges of automatically capturing experiment provenance, along with methods to store the resulting provenance data efficiently. REDUX uses the *Windows Workflow Foundation* (WinWF) as a workflow engine. The VisTrails system [27] represented an initial attempt to improve the scientific discovery process and reduce the time to insight using XML and relational database technologies for provenance management. VisTrails addressed the visualization problem from a data management perspective and managed the data and metadata of a visualization product. It can support scientists in navigating through the space of workflows and parameter settings for an exploration process.

The provenance models proposed by the semantic web community for data-driven workflows capture retrospective provenance but underspecify the workflow structure. An underspecified workflow structure may misinterpret scientific experiments and preclude the workflow's conformance checking, eventually restricting provenance. To overcome these challenges [67] proposed a formal lightweight and general purpose specification model for the control-flows involved workflows and integrated it with both ProvOne [66] and PROV-DM [90] provenance models.

Sometimes, large-scale experiments may demand a week to run, even in high-performance computing environments. So it becomes unviable to analyze provenance data only after the end of the execution. However, scientists can use run time provenance to monitor workflow execution and take action before execution end. For example, [6] worked with representing and sharing runtime provenance data to improve experiment management and analyze scientific data generated by parallel workflow execution in different environments adopting cartridges. Their works support runtime analysis, evaluate the status of each parallel task, take actions to improve workflow reliability and performance, spare financial resources, and steer the execution status at any time.

Apart from mentioned works CombeChem [36, 57], Mindswap [43, 44] and VIEW [31, 45] systems worked with provenance management. Swift [37], and Chimera [46]

systems introduce a Virtual Data System (*VDS*) to use provenance for tracking the data derivation history, on-demand data generation and re-generation, and data product validation. Chimera [46] combines a virtual data catalog for representing data derivation procedures and derived data with a virtual data language interpreter that translates user requests into data definition and query operations on the database. They coupled the Chimera system with distributed "data grid" services to enable on-demand execution of computation schedules constructed from database queries.

B. Our research in the context of related work

Several available SWfMSs use general-purpose relational RDF, structured files, relational tables, OWL schemas, or virtual collections to manage and query provenance. For this research, we planned to use a relational database to store provenance data and optimized SQL to query the provenance information. We mainly focused on two things.

1) elimination of redundancies based on the pattern and database schema semantics. This approach can be used in any general-purposed RDF stores or relational tables.
2) we also planned to store data, modules, and invocations information at a granular level to re-use it to ensure the re-usability of provenance data.

Our target is to provide a provenance management approach that eases the gathering, storage, and analysis of provenance information so that scientists can use provenance data without putting the burden on adaptations.

V. Study Design

This study aimed to investigate the extent to which SWfMSs and different provenance management systems work mechanisms and discovered the challenges of managing provenance data. We first explored state-of-the-art SWfMSs' provenance capturing, storing, and usability studies for this research. Then, we explored several provenance management models and the necessity of exposing provenance in open science (we can not provide comprehensive coverage of all systems due to space limitations; we review a representative set of widely used systems). Finally, we demonstrated the challenges and suggested possible actions to overcome them.

For practicing open science, Galaxy SWfMSs enabled users to share workflows with provenance data into several workflow repositories [92, 93, 94]. Galaxy has an internal proprietary provenance model and captures both prospective and retrospective provenance information; also, there are facilities for adding manual annotations (important decisions and notes, usually used to understand the meaning of data products or scientific applications). But Galaxy allows data redundancies, and the facilities for reusing the provenance data for workflow re-execution are unavailable. For example, if one user needs to execute a workflow multiple times, the current Galaxy system executes it from the beginning. Therefore, it cannot reuse the provenance information for re-execution. Details analysis is provided in the study finding section. We also go through the Galaxy repositories [92, 93, 94] and identified most of the workflows, as well as datasets, are redundant, which hinder SWfMSs' performance and cost enormous resources.

Vistrails' [27] provenance-management system provides infrastructure for data exploration and visualization. Using a relational database management system Vistrails uses an action-based provenance model to capture changes to parameter values and pipeline definitions. It supports backward and forwards chains of reasoning. The Vistrails interface allows scientists to query, interact and visualize the process history. Though it supports the flexible reuse of workflow pipelines, collaborative exploration, and a flexible annotation framework, it does not support the reusability of provenance data for workflow re-execution. In VisTrails, the change-based provenance model records information about modifications to a task, akin to a database transaction log. It also stores the redundant data and allows redundant operations. In Taverna, we noticed two classes of provenance, one describes workflow-related entities like services, workflows, and sub-workflows, and the other describes workflow executions. It used a graph model to capture provenance information. In Taverna, the processors can exchange data by reference, which allows, among other things, to reduce the amount of data the workflow enactor has to convey between services, thereby supporting data-intensive processes. Taverna stores provenance metadata using RDF/XML stores, and here, multiple annotations can coexist and be associated with the same resource.

In Kepler [22], the specification of a workflow instance must be saved to the provenance model every time the workflow is executed, along with runtime information which incurs high storage overheads and negatively impacts query performance. Karma provenance framework [79] stores both prospective and retrospective provenance using XML and relational databases. It supports dynamic workflows and explicitly models data products' derivation history. It provides a light-weight and scalable implementation to meet the core needs of recording and querying for these provenance graphs over hundreds of thousands of service invocations and data products but allow data redundancies.

Several approaches have been developed for capturing and modeling provenance. For example, some models used traditional filesystem, and others used relational database tables. In the filesystem, the advantage is that users do not need additional infrastructure to store provenance information. The relational database provides centralized, efficient storage that a group of users can share. But there are several issues with storing, accessing, and querying provenance data.

VI. Study Findings

Although provenance models differ in several ways, including their use of structures and storage strategies, they all share an essential type of information: processes and data dependencies. Provenance is relevant to a wide range of domains and applications, so it is crucial to identify the problem of systematically capturing and managing provenance for computational tasks. Without provenance, it is nearly impossible to reproduce and share results, validate results with

a different set of input data, understand the operation process and solve a complex problem collaboratively. Therefore, we conducted this research and gathered the following findings in this study.

A. Answering RQ1

RQ1: What are the provenance challenges in Scientific Workflow Management Systems? By doing this exploratory study, we identified the following challenges.

Data Redundencies: Redundancy of data means multiple copies of the same information spread over multiple locations in the same database. Data redundancies inflate the size of the database and create data inconsistencies. If data is redundant, then data maintenance becomes tedious and problematic. By exploring several SWfMSs and provenance models, we found that they all support redundant data storing in provenance.

Lack of Reusability: Data reusability lessens the response time, and reusable data allows first mover advantages. Due to the open science policy, many scientists make their data available to the community. Still, if a workflow/workflow module needs multiple re-execution, current provenance management systems do not reuse the previous data and make new execution which causes several problems like storage issues, data fetching and updating, and many more issues

Data Querying Mechanisms: The ability to query provenance efficiently helps knowledge reusability, which helps compare and understand differences between different tasks. Current provenance management systems mostly use XML, RDF, and relational databases for storing and querying data. Unfortunately, the data querying mechanism is not user-friendly in most SWfMSs. For querying provenance data, most SWfMSs use SQL, Prolog, and SPARQL, which can be awkward and complex to write.

Provenance Data Sharing: Scientific studies often require a heterogeneous set of data, and these data sets are collected by independent research over many years, which are not accompanied by rich enough semantic information. We noticed several workflow repositories where provenance information is shared, but they are difficult to interpret as there has no annotation or usage purpose, or explanation.

Provenance Overload: Sometimes, a workflow execution can take multiple steps, and a module may need multiple executions. Current SWfMSs store all executed data. In this case, the information stored for a single workflow can be extensive. Thus, provenance overload can be a problem for these systems.

B. Answering RQ2

RQ2: How can we overcome the provenance challenges? In SWfMSs, scientists need to work with a very high volume of data, and therefore, infrastructure for effectively and efficiently querying provenance data is an essential component of a provenance management system. In order to resolve the challenges scientists usually face for provenance, we identified the following measures.

Provenance Database Normalization: Structuring a relational database using normal forms reduces data redundancy and improves data integrity. Structuring provenance information into multiple layers enables normalized representation, which avoids storing redundant data. Current SWfMSs provenance systems contain vast amounts of redundant data. To get rid of it, we are suggesting using 3NF [98] or BCNF [100].

Enable Data Reusability: Current SWfMSs and their provenance management systems use provenance information primarily for reproducibility and audit-trail purposes. But data reusability can reduce the resource usage and time to perform an operation because it will avoid processing of same data multiple times.

Effective Data Sharing: Proper access to data is essential for an efficient, progressive, and self-correcting scientific ecosystem. Though open science encourages more open access to and use of data as it helps collaboration among teams and communities, there is still a significant lack of shared data in workflow communities. Also, we found redundant and erroneous shared data in several workflow repositories. Therefore, we are encouraging adequate data sharing in the community.

Along with the mentioned reasons, we also encouraged fine-grained data storing, an intuitive and interactive interface for provenance queries, a generic provenance model so that data can be shared among SWfMSs, and the usage of the optimized query to analyze provenance data.

C. Answering RQ3

RQ3: Can we use provenance data in real-time? One of the primary perspectives of our research is to ensure provenance data, more specifically retrospective provenance data reusability in scientific workflow execution. The FAIR principle also encouraged more open access to, and use of data to solve problems, and many researchers considered reusability the primary concern. If we can store provenance data using data normalization and share the data following FAIR guidelines, then we can ensure data reusability.

VII. Conclusion and Future Work

Data-centric computing is increasingly becoming essential for scientific analysis, and the FAIR principles have represented a meaningful way forward for open science datasets. Efficient data storing and re-usability are becoming inevitable components in SWfMS as workflow may become very large, or the same workflow needs to be executed several times. In this research, we identified provenance challenges; more specifically, we focused on optimized storing and reusing provenance data. We also described ways to mitigate these challenges. In the future, we will build a provenance data storing system that will be compatible with any SWfMS and ensure optimized storing and reusing of provenance data along with reproduction of results from the earlier executions, explaining unexpected results and efficient data sharing.

ACKNOWLEDGMENT

This research is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grants, and by an NSERC Collaborative Research and Training Experience (CREATE) grant, and by two Canada First Research Excellence Fund (CFREF) grants coordinated by the Global Institute for Food Security (GIFS) and the Global Institute for Water Security (GIWS).

REFERENCES

[1] Silva, R., Filgueira, R., Pietri, I., Jiang, M., Sakellariou, R. & Deelman, E. A characterization of workflow management systems for extreme-scale applications. *FGCS*. **75** pp. 228-238 (2017)

[2] Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*. **11**, 1-13 (2010)

[3] Liu, J., Pacitti, E., Valduriez, P. & Mattoso, M. A survey of data-intensive scientific workflow management. *Journal Of Grid Computing*. **13**, 457-493 (2015)

[4] Davidson, S. & Freire, J. Provenance and scientific workflows: challenges and opportunities. *Proceedings Of The 2008 ACM SIGMOD*. pp. 1345-1350 (2008)

[5] Simmhan, Y., Plale, B. & Gannon, D. A survey of data provenance in e-science. *ACM Sigmod Record*. **34**, 31-36 (2005)

[6] Costa, F., Silva, V., De Oliveira, D., Ocaña, K., Ogasawara, E., Dias, J. & Mattoso, M. Capturing and querying workflow runtime provenance with PROV: a practical approach. *Proceedings Of The Joint EDBT/ICDT 2013 Workshops*. pp. 282-289 (2013)

[7] Deelman, E., Vahi, K., Juve, G., Rynge, M., Callaghan, S., Maechling, P., Mayani, R., Chen, W., Da Silva, R., Livny, M. & Others Pegasus, a WMS for science automation. *FGCS*. **46** pp. 17-35 (2015)

[8] Mazaheri, M., Kiar, G. & Glatard, T. A Recommender System for Scientific Datasets and Analysis Pipelines. *2021 IEEE Workshop On Workflows In Support Of Large-Scale Science (WORKS)*. pp. 1-8 (2021)

[9] Al-Saadi, A., Ahn, D., Babuji, Y., Chard, K., Corbett, J., Hategan, M., Herbein, S., Jha, S., Laney, D., Merzky, A. & Others ExaWorks: Workflows for Exascale. *2021 IEEE WORKS*. pp. 50-57 (2021)

[10] Özsu, M. & Valduriez, P. Principles of distributed database systems. (Springer,1999)

[11] Maechling, P., Deelman, E., Zhao, L., Graves, R., Mehta, G., Gupta, N., Mehringer, J., Kesselman, C., Callaghan, S., Okaya, D. & Others SCEC CyberShake workflows—automating probabilistic seismic hazard analysis calculations. *Workflows For E-Science*. pp. 143-163 (2007)

[12] E. Southern California earthquake center. *Caltech. Dataset*. (2013)

[13] Silva, R., Casanova, H., Chard, K., Laney, D., Ahn, D., Jha, S., Goble, C., Ramakrishnan, L., Peterson, L., Enders, B. & Others Workflows community summit: Bringing the scientific workflows community together. *ArXiv Preprint ArXiv:2103.09181*. (2021)

[14] Livny, J., Teonadi, H., Livny, M. & Waldor, M. High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs. *PloS One*. **3**, e3197 (2008)

[15] Abramovici, A., Althouse, W., Drever, R., Gürsel, Y., Kawamura, S., Raab, F., Shoemaker, D., Sievers, L., Spero, R., Thorne, K. & Others LIGO: The laser interferometer gravitational-wave observatory. *Science*. pp. 325-333 (1992)

[16] Taylor, I., Shields, M., Wang, I. & Harrison, A. The triana workflow environment: Architecture and applications. *Workflows For E-Science*. pp. 320-339 (2007)

[17] Kumar, A., Rasche, H., Grüning, B. & Backofen, R. Tool recommender system in Galaxy using deep learning. *GigaScience*. **10**, giaa152 (2021)

[18] Koop, D., Scheidegger, C., Callahan, S., Freire, J. & Silva, C. Viscomplete: Automating suggestions for visualization pipelines. *IEEE Trans. Vis. Comput. Graph.* **14**, 1691-1698 (2008)

[19] Soomro, K., Munir, K. & McClatchey, R. Incorporating semantics in pattern-based scientific workflow recommender systems: Improving the accuracy of recommendations. *2015 Science And Information Conference (SAI)*. pp. 565-571 (2015)

[20] Hossain, M., Roy, B., Roy, C. & Schneider, K. VizSciFlow: A Visually Guided Scripting Framework for Supporting Complex Scientific Data Analysis. *Proceedings Of The ACM On HCI*. **4**, 1-37 (2020)

[21] Foster, I., Vöckler, J., Wilde, M. & Zhao, Y. The Virtual Data Grid: A New Model and Architecture for Data-Intensive Collaboration.. *CIDR*. (2003)

[22] Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B. & Mock, S. Kepler: an extensible system for design and execution of scientific workflows. *Proceedings. 16th International Conference On Scientific And Statistical Database Management, 2004.*. pp. 423-424 (2004)

[23] Ludäscher, B., Weske, M., McPhillips, T. & Bowers, S. Scientific workflows: Business as usual?. *International Conference On Business Process Management*. pp. 31-47 (2009)

[24] Lim, C., Lu, S., Chebotko, A. & Fotouhi, F. Prospective and retrospective provenance collection in scientific workflow environments. *2010 IEEE International Conference On Services Computing*. pp. 449-456 (2010)

[25] Holland, D., Seltzer, M., Braun, U. & Muniswamy-Reddy, K. PASSing the provenance challenge. *Concurrency And Computation: Practice And Experience*. **20**, 531-540 (2008)

[26] Barga, R. & Digiampietri, L. Automatic capture and efficient storage of e-Science experiment provenance. *Concurrency And Computation: Practice And Experience*. **20**, 419-429 (2008)

[27] Callahan, S., Freire, J., Santos, E., Scheidegger, C., Silva, C. & Vo, H. VisTrails: visualization meets data management. *Proceedings Of The 2006 ACM SIGMOD International Conference On Management Of Data*. pp. 745-747 (2006)

[28] Agrawal, R., Imieliński, T. & Swami, A. Mining association rules between sets of items in large databases. *Proceedings Of The 1993 ACM SIGMOD*. pp. 207-216 (1993)

[29] Junaid, M., Berger, M., Vitvar, T., Plankensteiner, K. & Fahringer, T. Workflow composition through design suggestions using design-time provenance information. *2009 5th IEEE Int. Conf. On E-Science Workshops*. pp. 110-117 (2009)

[30] Chakroborti, D., Mondal, M., Roy, B., Roy, C. & Schneider, K. Optimized Storing of Workflow Outputs through Mining Association Rules. *2018 IEEE Big Data*. pp. 508-515 (2018)

[31] Chebotko, A., Fei, X., Lu, S. & Fotouhi, F. Scientific workflow provenance metadata management using an RDBMS-based RDF store. *Wayne State University, Tech. Rep. TR-DB-092007-CFLF*. (2007)

[32] Graves, R., Jordan, T., Callaghan, S., Deelman, E., Field, E., Juve, G., Kesselman, C., Maechling, P., Mehta, G., Milner, K. & Others CyberShake: A physics-based seismic hazard model for southern California. *Pure And Applied Geophysics*. **168**, 367-381 (2011)

[33] Chakroborti, D., Roy, B. & Nath, S. Designing for Recommending Intermediate States in A Scientific Workflow Management System. *Proceedings Of The ACM On HCI*. **5**, 1-29 (2021)

[34] Anand, M., Bowers, S. & Ludäscher, B. Techniques for efficiently querying scientific workflow provenance graphs.. *EDBT*. **10**, 287-298 (2010)

[35] Frew, J., Metzger, D. & Slaughter, P. Automatic capture and reconstruction of computational provenance. *Concurrency And Computation: Practice And Experience*. **20**, 485-496 (2008)

[36] Taylor, K., Gledhill, R., Essex, J., Frey, J., Harris, S. & De Roure, D. Bringing chemical data onto the semantic web. *Journal Of Chemical Information And Modeling*. **46**, 939-952 (2006)

[37] Zhao, Y., Hategan, M., Clifford, B., Foster, I., Von Laszewski, G., Nefedova, V., Raicu, I., Stef-Praun, T. & Wilde, M. Swift: Fast, reliable, loosely coupled parallel computation. *2007 IEEE Congress On Services (Services 2007)*. pp. 199-206 (2007)

[38] De Oliveira, D., Ogasawara, E., Baião, F. & Mattoso, M. Scicumulus: A lightweight cloud middleware to explore many task computing paradigm in scientific workflows. *2010 IEEE 3rd CLOUD*. pp. 378-385 (2010)

[39] Mattoso, M., Werner, C., Travassos, G., Braganholo, V., Ogasawara, E., Oliveira, D., Cruz, S., Martinho, W. & Murta, L. Towards supporting the life cycle of large scale scientific experiments. *International Journal Of Business Process Integration And Management*. **5**, 79 (2010)

[40] De Oliveira, D., Ocana, K., Ogasawara, E., Dias, J., Gonçalves, J., Baiao, F. & Mattoso, M. Performance evaluation of parallel strategies in public clouds: A study with phylogenomic workflows. *FGCS*. **29**, 1816-1825 (2013)

[41] Ocaña, K., Oliveira, D., Dias, J., Ogasawara, E. & Mattoso, M. Discovering drug targets for neglected diseases using a pharmacophylogenomic cloud workflow. *2012 IEEE 8th E-Science*. pp. 1-8 (2012)

[42] Ocaña, K., Oliveira, D., Ogasawara, E., Dávila, A., Lima, A. & Mattoso, M. SciPhy: a cloud-based workflow for phylogenetic analysis of drug targets in protozoan genomes. *BSB*. pp. 66-70 (2011)

[43] Golbeck, J. Combining provenance with trust in social networks for semantic web content filtering. *IPAW*. pp. 101-108 (2006)

[44] Golbeck, J. & Hendler, J. A semantic web approach to the provenance challenge. *Concurrency And Computation: Practice And Experience*. **20**,

431-439 (2008)
[45] Lin, C., Lu, S., Lai, Z., Chebotko, A., Fei, X., Hua, J. & Fotouhi, F. Service-oriented architecture for VIEW: a visual scientific workflow management system. *2008 IEEE SCC*. **1** pp. 335-342 (2008)
[46] Foster, I., Vockler, J., Wilde, M. & Zhao, Y. Chimera: A virtual data system for representing, querying, and automating data derivation. *Proceedings 14th SSDBM*. pp. 37-46 (2002)
[47] Galaxy SWFMS, https://usegalaxy.org/, Online; Last accessed July 2022
[48] Vistrails SWFMS, https://www.vistrails.org/index.php/, Online; Last accessed July 2022
[49] Goff, S., Vaughn, M., McKay, S., Lyons, E., Stapleton, A., Gessler, D., Matasci, N., Wang, L., Hanlon, M., Lenards, A. & Others The iPlant collaborative: cyberinfrastructure for plant biology. *Frontiers In Plant Science*. **2** pp. 34 (2011)
[50] Cruz, S., Barros, P., Bisch, P., Campos, M. & Mattoso, M. Provenance services for distributed workflows. *2008 Eighth IEEE International Symposium On CCGRID*. pp. 526-533 (2008)
[51] Mattoso, M. & Others Management of Scientific Experiments in Large Scale. *XXVIII SBC Conference,–Belém, PA, Brazil. Http://gexp. Nacad. Ufrj. Br/documents*. (2008)
[52] Deelman, E., Gannon, D., Shields, M. & Taylor, I. Workflows and e-Science: An overview of workflow system features and capabilities. *FGCS*. **25**, 528-540 (2009)
[53] Groth, P., Miles, S., Fang, W., Wong, S., Zauner, K. & Moreau, L. Recording and using provenance in a protein compressibility experiment. *HPDC-14. Proceedings.2005.*. pp. 201-208 (2005)
[54] Buneman, P., Chapman, A. & Cheney, J. Provenance management in curated databases. *Proceedings Of The 2006 ACM SIGMOD*. pp. 539-550 (2006)
[55] Kim, J., Deelman, E., Gil, Y., Mehta, G. & Ratnakar, V. Provenance trails in the wings/pegasus system. *Concurrency And Computation: Practice And Experience*. **20**, 587-597 (2008)
[56] Groth, P., Jiang, S., Miles, S., Munroe, S., Tan, V., Tsasakou, S. & Moreau, L. An architecture for provenance systems. (2006)
[57] Frey, J., De Roure, D., Taylor, K., Essex, J., Mills, H. & Zaluska, E. CombeChem: a case study in provenance and annotation using the Semantic Web. *IPAW*. pp. 270-277 (2006)
[58] Bose, R. & Frew, J. Lineage retrieval for scientific data processing: a survey. *ACM Computing Surveys (CSUR)*. **37**, 1-28 (2005)
[59] Holland, D., Braun, U., Maclean, D., Muniswamy-Reddy, K. & Seltzer, M. Choosing a data model and query language for provenance. *Proceedings Of The 2nd IPAW'08*. (2008)
[60] Dias, J., Guerra, G., Rochinha, F., Coutinho, A., Valduriez, P. & Mattoso, M. Data-centric iteration in dynamic workflows. *FGCS*. **46** pp. 114-126 (2015)
[61] Berriman, G., Deelman, E., Good, J., Jacob, J., Katz, D., Kesselman, C., Laity, A., Prince, T., Singh, G. & Su, M. Montage: a grid-enabled engine for delivering custom science-grade mosaics on demand. *Optimizing Scientific Return For Astronomy Through Information Technologies*. **5493** pp. 221-232 (2004)
[62] Juve, G., Chervenak, A., Deelman, E., Bharathi, S., Mehta, G. & Vahi, K. Characterizing and profiling scientific workflows. *FGCS*. **29**, 682-692 (2013)
[63] Ocaña, K., Oliveira, D., Horta, F., Dias, J., Ogasawara, E. & Mattoso, M. Exploring molecular evolution reconstruction using a parallel cloud based scientific workflow. *BSB*. pp. 179-191 (2012)
[64] Fahringer, T., Prodan, R., Duan, R., Hofer, J., Nadeem, F., Nerieri, F., Podlipnig, S., Qin, J., Siddiqui, M., Truong, H. & Others Askalon: A development and grid computing environment for scientific workflows. *Workflows For E-Science*. pp. 450-471 (2007)
[65] Sahoo, S., Sheth, A. & Henson, C. Semantic provenance for escience: Managing the deluge of scientific data. *IEEE Internet Computing*. **12**, 46-54 (2008)
[66] Curcin, V., Ghanem, M., Wendel, P. & Guo, Y. Heterogeneous workflows in scientific workflow systems. *ICCS*. pp. 204-211 (2007)
[67] Butt, A. & Fitch, P. Provone+: a provenance model for scientific workflows. *ICCS Engineering*. pp. 431-444 (2020)
[68] Altintas, I., Barney, O. & Jaeger-Frank, E. Provenance collection support in the kepler scientific workflow system. *IPAW*. pp. 118-132 (2006)
[69] Marinho, A., Murta, L., Werner, C., Braganholo, V., Cruz, S., Ogasawara, E. & Mattoso, M. ProvManager: a provenance management system for scientific workflows. *Concurrency And Computation: Practice And Experience*. **24**, 1513-1530 (2012)
[70] Cruz, S., Campos, M. & Mattoso, M. Towards a taxonomy of provenance in scientific workflow management systems. *2009 Congress On Services-I*. pp. 259-266 (2009)
[71] Zhao, J., Goble, C., Stevens, R. & Turi, D. Mining Taverna's semantic web of provenance. *Concurrency And Computation: Practice And Experience*. **20**, 463-472 (2008)
[72] Belhajjame, K., Wolstencroft, K., Corcho, O., Oinn, T., Tanoh, F., William, A. & Goble, C. Metadata management in the taverna workflow system. *2008 Eighth IEEE International Symposium On Cluster Computing And The Grid (CCGRID)*. pp. 651-656 (2008)
[73] Srivastava, D. & Velegrakis, Y. Intensional associations between data and metadata. *Proceedings Of The ACM SIGMOD*. pp. 401-412 (2007)
[74] Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., Silva Santos, L., Bourne, P. & Others The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. **3**, 1-9 (2016)
[75] Missier, P., Soiland-Reyes, S., Owen, S., Tan, W., Nenadic, A., Dunlop, I., Williams, A., Oinn, T. & Goble, C. Taverna, reloaded. *SSDBM*. pp. 471-481 (2010)
[76] Balaskó, Á. Workflow concept of ws-pgrade/guse. *Science Gateways For Distributed Computing Infrastructures*. pp. 33-50 (2014)
[77] Zhao, J., Wroe, C., Goble, C., Stevens, R., Quan, D. & Greenwood, M. Using semantic web technologies for representing e-science provenance. *ISWC*. pp. 92-106 (2004)
[78] Buneman, P. & Tan, W. Provenance in databases. *Proceedings Of The 2007 ACM SIGMOD International Conference On Management Of Data*. pp. 1171-1173 (2007)
[79] Simmhan, Y., Plale, B. & Gannon, D. A framework for collecting provenance in data-centric scientific workflows. *2006 IEEE International Conference On Web Services (ICWS'06)*. pp. 427-436 (2006)
[80] Bharathi, S., Chervenak, A., Deelman, E., Mehta, G., Su, M. & Vahi, K. Characterization of scientific workflows. *2008 Third Workshop On Workflows In Support Of Large-scale Science*. pp. 1-10 (2008)
[81] Park, S., Faraci, G., Ward, P., Emerson, J. & Lee, H. High-precision and cost-efficient sequencing for real-time COVID-19 surveillance. *Scientific Reports*. **11**, 1-10 (2021)
[82] Ogasawara, E., De Oliveira, D., Valduriez, P., Dias, J., Porto, F. & Mattoso, M. An algebraic approach for data-centric scientific workflows. *Proceedings Of The VLDB Endowment (PVLDB)*. **4**, 1328-1339 (2011)
[83] Wozniak, J., Jain, R., Balaprakash, P., Ozik, J., Collier, N., Bauer, J., Xia, F., Brettin, T., Stevens, R., Mohd-Yusof, J. & Others Candle/supervisor: A workflow framework for machine learning applied to cancer research. *BMC Bioinformatics*. **19**, 59-69 (2018)
[84] Sivaraman, G., Jackson, N., Sanchez-Lengeling, B., Vázquez-Mayagoitia, Á., Aspuru-Guzik, A., Vishwanath, V. & De Pablo, J. A machine learning workflow for molecular analysis: application to melting points. *MLST*. **1**, 025015 (2020)
[85] Galaxy Main Repo., https://usegalaxy.org/workflows/list_published, Online; Last accessed July 2022
[86] WorkflowHub, https://workflowhub.eu/workflows, Online; Last accessed July 2022
[87] Galaxy EU Repo., https://usegalaxy.eu/workflows/list_published, Online; Last accessed July 2022
[88] Galaxy AU Repo., https://usegalaxy.org.au/workflows/list_published, Online; Last accessed July 2022
[89] myExperiment, https://www.myexperiment.org/workflows, Online; Last accessed July 2022
[90] Belhajjame, K., B'Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T., McCusker, J. & Others Prov-dm: The prov data model. *W3C Recommendation*. **14** pp. 15-16 (2013)
[91] Dockstore, https://dockstore.org/organizations, Last accessed July 2022
[92] Galaxy Main Shared Histories,https://usegalaxy.org/histories/list_published, Last accessed July 2022
[93] Galaxy EU Shared Histories,https://usegalaxy.eu/histories/list_published, Last accessed July 2022
[94] Galaxy EU Shared Histories,https://usegalaxy.org.au/histories/list_published, Last accessed July 2022
[95] Oliveira, W., Oliveira, D. & Braganholo, V. Provenance analytics for workflow-based computational experiments: A survey. *ACM Computing Surveys (CSUR)*. **51**, 1-25 (2018)
[96] Prabhune, A., Zweig, A., Stotzka, R., Hesser, J. & Gertz, M. P-PIF: a ProvONE provenance interoperability framework for analyzing heterogeneous workflow specifications and provenance traces. *Distributed And Parallel Databases*. **36**, 219-264 (2018)
[97] Pérez, J., Arenas, M. & Gutierrez, C. Semantics and complexity of

SPARQL. *ACM Transactions On Database Systems (TODS)*. **34**, 1-45 (2009)
[98] Codd, E. Further normalization of the data base relational model. *Data Base Systems*. **6** pp. 33-64 (1972)
[99] Butt, A., Car, N. & Fitch, P. Towards Ontology Driven Provenance in Scientific Workflow Engine.. *MODELSWARD*. pp. 105-115 (2020)
[100] Codd, E. Recent Investigations in Relational Data Base Systems.. (1975)