university of groningen

faculty of science and engineering

Provenance of Adaptation in Scientific workflows - Literature review

Bachelor's Project Computer Science

August 2024

Author: Julia Dahlberg
Student Number: S4475151
First supervisor: Dimka Karastoyanova
Second supervisor: Ludwig Stage

Abstract

In the world of science new technology have opened up the possibility to rely on advanced computational methods and models to conduct and produce scientific research. An important aspect of scientific and business workflows is provenance - which refers to the information describing the production or history of an end product. This end product can be anything from digital data to a physical object, but in this context its usually referred to describe digital data. Capturing provenance is thereby also important when working with workflows, and although there are many different tools and systems to be used for this purpose, the challenge of capturing and analyzing provenance information about adaptive workflows persists. To address this issue we will examine the capture of provenance information in scientific and business workflows, as well as provenance of adaptations, and how to utilize and visualize this type of information. This is necessary to get a better understanding of how to recreate an adaptive workflow, as well as to get a clear overview of the topic of provenance and to realize what information might be lacking.

Contents

List of Figures

List of Tables

1 Introduction

The technology is ever so quickly evolving in our society, which opens up new possibilities in many different areas but also pushes us to adapt to the new innovations. One of the areas affected by this change is science. New technology have opened up the possibility to rely on advanced computational methods and models to conduct and produce scientific research. With this new way of doing research new findings are possible, but new challenges are also presented. One of these challenges is presented in the aspect of reproducible research [41, 26]. For in silico research to be deemed reproducible, there needs to be detailed information accessible about the software environment used, which data was used and produced and how every step was carried out. All of this information supports establishing the provenance of the research.

Provenance is not only crucial in scientific experiments but also in business workflows. In scientific and business workflows provenance describes the production or history of a data product. The capturing of provenance is a topic that has gained importance, at least since the increase of in silico research and simulated experiments with its handling of large amounts of data and additional levels of complexity. As a result of this there are different tools and systems to be used for capturing and managing provenance in scientific and business workflows.

There are many different aspects of provenance and one challenge is how to capture them all, or alternatively how to decide what aspects are needed to ensure reproducibility. One of these aspects that we will examine in this paper is the information of change or adaptation in a workflow during runtime. This change may occur for example when a scientist decides to change the software used for a part of the workflow, or the dataset used receives some additional data-points. These workflow changes are crucial for the trial-and-error manner of conducting their experiments and their research, making it important to capture this information. As of today, many of these important details of the process might get lost, which in turn would affect the reproducibility aspect of the experiment.

In this paper we report the findings made in a systematic literature review on this very topic. We will examine different methods and tools used to capture and visualize provenance, as well as what is missing in this area. With this information and potential future solutions we will identify future challenges, research topics, and open questions for investigation.

1.1 Provenance

Within scientific research documentation is of great importance. Clear documentation ensures that all aspects of the research process, including methodologies, data collection procedures, and analytical techniques, are clearly established. This transparency is essential for others to replicate the research and verify its findings, thereby promoting reproducibility and reliability in scientific inquiry.

This has arisen as one of the challenges of eScience, as reproducing an advanced computational experiment needs a lot of specific data and methods which can be difficult to

document and keep track of during the research process. This is where the concept of provenance information becomes a crucial aspect [41, 26]. As shown through the process of this review, provenance can be defined in different ways often depending on the intended usage. A general definition however could be summarized as the history or detailed story of how some final product was derived, in this case the final product of a workflow.

1.1.1 Provenance Taxonomy

There are different ways to define the term provenance, although all definitions are related to describe the origin or history of some form. As stated in [30] it is often dependent on the field in which the term is used. For example within the scientific community it is often used to describe a scientific workflow, but it could also be used to focus more on the insights and hypotheses.

In this report, we base our definition of provenance on the work by Herschel et al. [17]. This definition serves as a foundation for exploring the different features of provenance information. One of these features is the granularity, which indicates the amount of detail in the provenance. There are two different granularities identified according to [17]: fine-grained and coarse-grained granularity. This feature may be mentioned but is not a primary focus within this work.

Another feature is the provenance-type. There are four different types of provenance depending on the area of use: provenance meta-data, information systems provenance, workflow provenance and data provenance [17]. In this paper we are going to focus on scientific- and business workflows. Therefore the provenance type of focus in this paper is going to be workflow provenance.

The last feature of provenance, and the one we will examine the most in this report, is the provenance-form. In [17] there were three different provenance-forms identified: prospective, retrospective and evolution provenance.

Prospective is the provenance that describes the structure or static context of a workflow, which means that it is not dependent on the input or execution of the workflow. Retrospective provenance on the other hand contains the specific execution information of a workflow. This includes information of the execution of every workflow step and the environment, as well as the accessed or produced resources. The last form, evolution provenance describes the changes between two different versions of a workflow.

In [4] the authors provide a clear and effective demonstration of these various provenance forms using the process of building an IKEA table as an example. They explain that the prospective provenance information acts as the manual for the table, with a clear plan or instruction on how to produce the desired product. Retrospective provenance on the other hand is the implementation of this plan, as some steps may not be executed precisely as intended. For example if a specific tool included in the manual is not accessible, an alternative may be used instead. As such, if a screw is missing one could use glue as a replacement. These types of information about how the process was actually implemented is described by retrospective provenance information. Evolution provenance was not mentioned in [4], but with the same analogy evolution provenance could be described as the differences between two versions of the IKEA table. For example they could produce a newer version of the table with more durable screws and a higher quality paint. Then the evolution provenance would

describe these new changes between the two tables.

In this report we are also going to examine the state of provenance of adaptation. This term is used by [33] to describe the adaptation or change in a workflow, more specifically ad-hoc changes and changes during runtime. Provenance of adaptation is thereby necessary to enable provenance of adaptive workflows. Provenance of change has been subdivided in [33] into the forms evolution provenance (originally mentioned in [17]) and ad-hoc provenance. The provenance taxonomy as defined by [17], and with the addition of [33] can be seen in Figure 1.

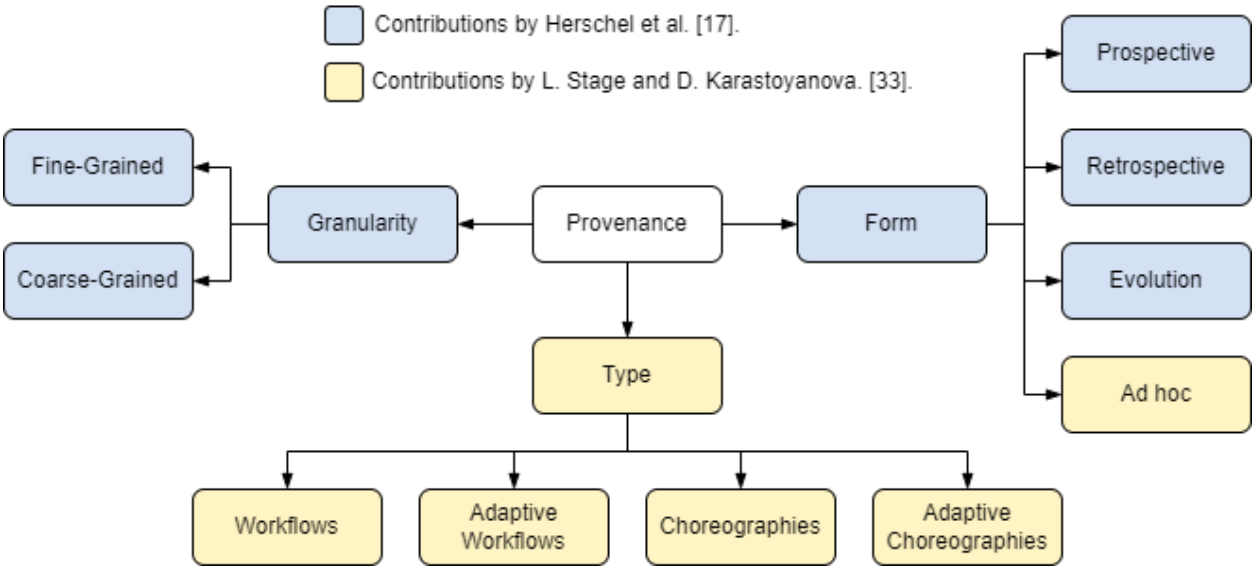

Figure 1: Workflow Provenance types taxonomy [17, 33]

1.1.2 Provenance Capturing (tools)

Depending on the use case, different data sources, processing pipelines and parties involved, the capturing and managing of provenance can look very different. As there are also currently no adopted global standards for provenance, the specific data collected by a provenance capture tool can vary greatly. The capture can be done either by the Workflow Management System (WfMS), a separate tool connected to the WfMS, or a standalone tool. The specific data actually captured by either method also varies. Some tools may only collect partial provenance, the types and forms of provenance can vary or if the data collected is fine-grained or coarse-grained.

1.1.3 Provenance Visualization

The visual representation of the provenance information can also vary between different approaches, as there is not a given standard. Visualization can be anything from a simple log in a document to complex and in-depth graphs.

1.2 Business and Scientific workflows

Business workflows are defined by [14] as a description of tasks, only specifically within a business process. These tasks are described on a theoretical level to enable the understanding of the business process, as well as for evaluating, and redesigning the process. The business workflow may also include information describing the process requirements [22, 40].

A scientific workflow shares many similarities with a business process, and can even be modeled as one. According to [24] the scientific workflow is defined as follows: a description of a scientific process, breaking it down into a series of smaller steps. It is constructed of components called "tasks" and "dependencies". The tasks represent computational steps used in for example data analysis and simulations, and are usually arranged and executed in a specific order. The dependencies represent the relationships between these tasks in the workflow. For example the output of one task might be required as the input of another task, so tasks have the possibility of being dependent on each other.

2 Methodology

The goal with this paper is to examine the current state of provenance in scientific and business workflows, and more specifically provenance of adaption. For this purpose we have formulated a selection of research questions. To explore this topic and answer the research questions the chosen methodology is the systematic literature review approach. This methodology was chosen as it is deemed an appropriate way to comprehensively review research on a specific topic, and as there was a limited amount of papers written on this topic it was also feasible to perform by one person. The review approach will follow the guidelines of [21], and is divided into multiple different steps that will be described further in this chapter.

2.1 Research questions

To explore the topic of provenance of adaptation in scientific and business workflows and how to capture it we have identified one main research question (RQ1) with five sub-questions (cf. Table 1). This approach allows us to systematically analyze various facets of the topic, while ensuring a comprehensive and in-depth investigation.

Table 1: Research questions

RQ1	How do we capture provenance information of adaptive scientific and business workflows?
RQ2	What information should be captured, and what is currently being captured?
RQ3	How is such information being used?
RQ4	How is such information being visualized?
RQ5	What are the available works and tools for this purpose?
RQ6	What is currently missing within this subject?

By addressing these questions, we aim to develop a deep understanding of not only how to capture provenance but also how to effectively use and visualize this data. The comparison between the ideal and current practices (RQ2), coupled with the evaluation of existing tools (RQ5), will highlight areas of strength and opportunities for improvement. The practical applications (RQ3) and visualization techniques (RQ4) will provide insights into the usability and accessibility of provenance, ensuring that it serves its intended purpose. Finally, by identifying what is missing in the current landscape (RQ6), we can propose new directions for research and development, ultimately contributing to more robust and efficient workflow management systems in both scientific and business contexts.

2.2 Scope

The first step in this process was to identify the scope. The topic was defined as provenance of adaptation in regards to scientific and business workflows. Specifically, we will examine the

workflow-provenance type and explore all forms and levels of granularity of provenance that we encounter. We will include papers on provenance in scientific- and business workflows, encompassing both those that involve adaptations and those that do not, as these can be used to answer the sub-questions and still provide valuable insight to the main research question. We will include papers related to scientific and business workflows - regardless of the specific field of study. We will also include papers on business processes and scientific choreographies shall we encounter this.

Our goal is to investigate the current state of the art of this topic, using [17] as a foundation. Therefore, we aim to use papers published subsequent to the release of this report in this review, or potentially older studies they may have overlooked or excluded for any reason.

2.3 Literature search

With the scope and research-questions identified, the next step in the systematic review process was to perform the literature search. This is the step were we define the search terms and perform our search on each of the selected search sources. The goal is to find and collect every paper that is related to our topic and can be used to answer the research-questions. The search terms and search sources are documented to confirm that the literature search is reproducible, as this is an important aspect of the research. The search terms and search sources used are listed in Table 2 and Table 3 respectively. As search term 2 were a more broadly defined search term meant to include papers not related to adaptation, this term were used for search in the title or abstract of the papers.

Table 2: Search terms for provenance in workflows

Query	Description
1.	provenance AND (adaptive OR adaptation) AND ((scientific OR business) AND workflow)
2.	provenance AND (scientific workflows OR business workflows OR adaptive workflows)
3.	provenance AND (adaptive workflows OR flexible workflows)

Table 3: Search sources for literature review

Source	Description
1	Web of Science [2]
2	Scopus [1]
3	IEEE Xplore [18]
4	DBLP (Digital Bibliography & Library Project) [12]

After defining the search-terms and sources the literature search was performed. During this process, as shown in Figure 2, all the results was initially judged based on their title

and the abstract to get an initial perception whether the paper was relevant or not to the topic. All papers that were deemed relevant were then collected and proceeded to the next step. Some studies were also gathered from the reference lists of those identified through the literature search.

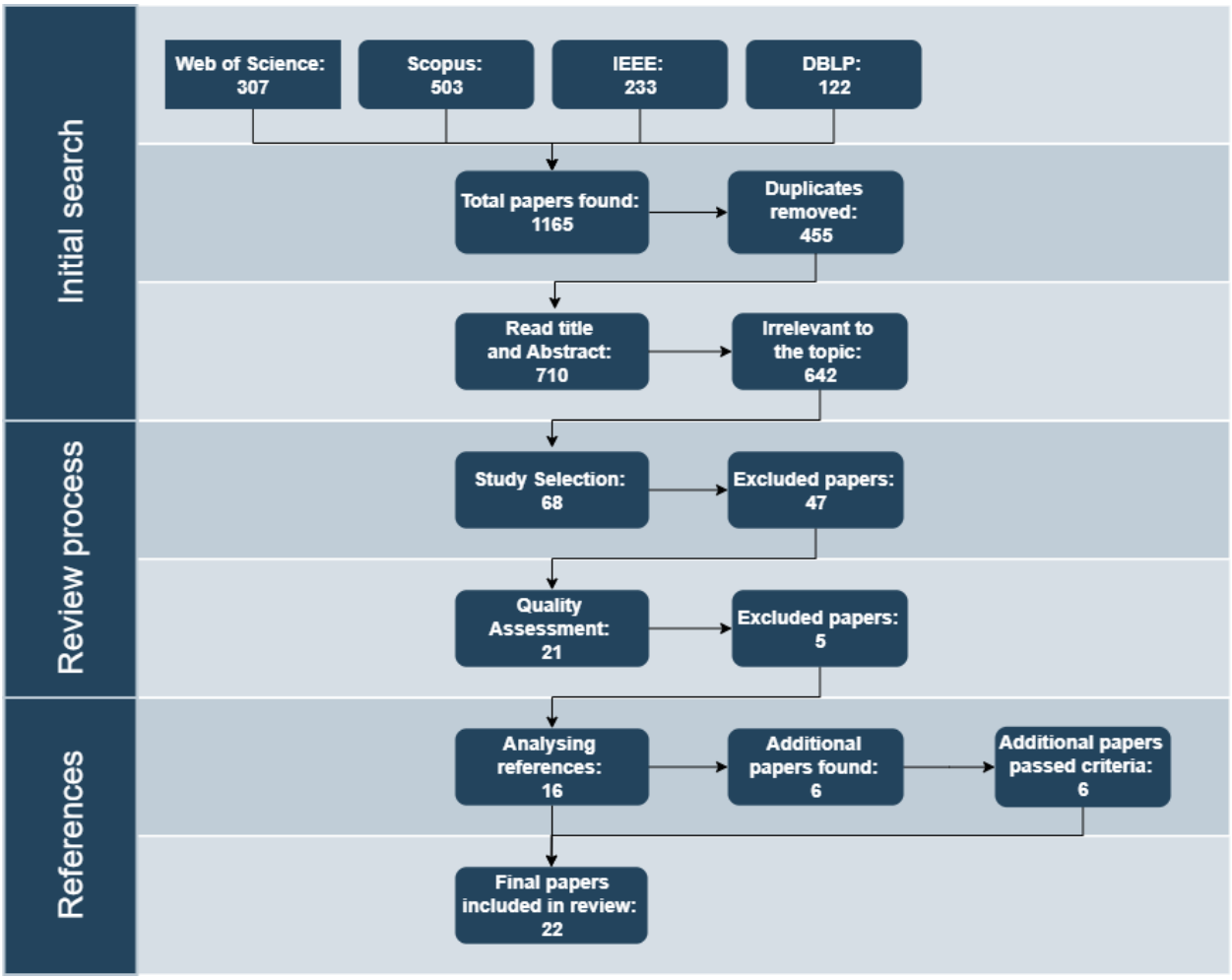

Figure 2: Literature Search Process and Iterations

2.4 Review protocol

Once the initial literature search has been completed, the next step is to decide which potential studies to include in the review. This is illustrated as the *Review process* in Figure 2. The selection process requires a systematic approach to ensure that only relevant and high-quality studies are considered. To facilitate this, a review protocol is necessary. The review protocol serves as a comprehensive guide that aids in both the selection of studies and the review process itself. It outlines the criteria for inclusion and exclusion, specifies the quality assessment, and details the methods for data extraction and analysis. By adhering to a well-defined protocol, the review process becomes more transparent, reproducible, and free

from bias, ensuring that the final review is both rigorous and reliable.

2.4.1 Study Selection

To select these studies a selection criteria is used to assess the studies' actual relevance. The selection criteria also help limit the chance of bias. There are two kinds of selection criteria: inclusion- and exclusion criteria. The former describes papers to be included in the review, and the latter describes papers to be excluded. The selection criteria used in this review are stated in Table 4 and Table 5.

Table 4: Inclusion criteria

Criterion	**Details**
Relevance to the topic	Papers must include the terms 'provenance', 'workflow', and preferably 'adaptive'.
Publication type	Peer-reviewed journal articles, conference papers, and academic theses are to be included.
Language	Papers written in English are included for accessibility and comprehensibility.
Scope	Papers covering different aspects such as methods, tools, techniques, challenges, or best practices related to provenance in workflows, preferably adaptive, are included.

Table 5: Exclusion Criteria

Criterion	**Details**
Lack of Detail	Papers lacking sufficient detail about relevant topics are excluded.

2.4.2 Quality Assessment

Once a paper has been selected using the selection criteria, it is important to evaluate the quality of the paper in question before proceeding to the review step. To perform this evaluation, a quality assessment is necessary. This process serves several important functions. First of all it ensures that the studies included in the review are both credible and reliable. It also helps to prevent bias as studies with weak methodologies or poor execution can introduce bias into the review, resulting in misleading results and conclusions. Apart from this purpose, the quality assessment can also serve as an additional selection criteria, as a study may meet the initial selection criteria but lack in quality. In cases like this the paper should be excluded from the review based on the quality assessment results. This process ensures that only the most reliable studies are included in the final review. The quality criteria used in this review are illustrated in Table 6.

Table 6: Quality Assessment

Criterion	**Details**
Clarity of research objective	• Are the research objectives clearly stated and well-defined? • Is it clear how the study contributes to understanding provenance in (adaptive) workflows?
Methodological rigor	• Is the methodology used appropriate for addressing the research questions? • Are potential biases or limitations of the methodology acknowledged?
Relevance and novelty	• Is the research situated within the context of existing literature, and does it build upon previous work effectively?
Results Interpretation	• Are the results clearly presented and interpreted? • Do the conclusions logically follow from the results presented? • Are limitations and potential sources of error adequately discussed?

The quality assessment for this review ensures that the studies presents a clear research objective, and uses an appropriate methodology for addressing the research objective. The studies should also demonstrate relevance and novelty, as well as a clear presentation and effective understanding of the results.

2.4.3 Data extraction

After selecting the studies to be included in the review using the selection criteria and the quality assessment we proceed to the data extraction. This is the step were the reviewer(s) reads through the selected studies and documents the relevant data from each study that may be utilized to answer the research-questions. To ensure consistency while reviewing and to guarantee reproducibility in this step, a data extraction form is constructed. The data extraction form serves as a tool to record and collect all the necessary information from the studies to answer the research questions. The same data extraction form is used for each paper reviewed. The template form used for this review can be seen in Table 7.

The *General Information* section collects the basic bibliographic details of each study. These details help in tracking and referencing each study, providing context regarding its academic or professional origins. In *Study Characteristics* the main objective and focus of the study are outlined. It identifies the type of workflow the study examines (if any) and whether it discusses the topic of adaptation in workflows or provenance. This information helps categorize the study and align it with the relevant research questions. In the third section *Provenance Capturing* we start to record more of the relevant data, as this section

Table 7: Data extraction form

Category	Details
General Information	• **Study ID:** Unique identifier for the study. • **Title:** Full title of the paper. • **Authors:** List of authors. • **Publication Year:** Year the study was published. • **Source:** Journal name or conference where the study was published. • **Type of Study:** (e.g., empirical study, conceptual study etc).
Study Characteristics	• **Study Objective:** What is the main objective of the study? • **Workflow Type:** What type of workflow does the study focus on? (e.g., adaptive workflows, business workflows, scientific workflows) • **Adaptation:** Does this paper discuss the topic of adaptation in workflows/provenance? (Yes/No)
Provenance Capturing	• **Provenance Definition:** How does the study define provenance? • **Provenance Data Captured:** What specific types of provenance are captured? (Granularity, type, form, etc.). • **Tools and Methods:** Which tools or methods are used to capture provenance? (e.g., specific software, frameworks).
Provenance Utilization	• **Use Cases:** How is the captured provenance utilized? (e.g., workflow recreation, debugging, optimization, auditing). • **Challenges:** What challenges are identified in capturing or utilizing provenance information?
Provenance Visualization	• **Visualization Techniques:** What techniques are used to visualize provenance information? (e.g., graphs, text-logs). • **Tools for Visualization:** What tools/methods are used for visualization?
Gaps and Future Work	• **Identified Gaps:** What gaps or missing elements are identified?
Quality Assessment	• **Study Design:** Is the study design appropriate to address the research question(s)? (Yes/No) • **Methodology:** Is the methodology well-defined and appropriate? (Yes/No) • **Bias and Limitations:** Are potential biases and limitations discussed? (Yes/No) • **Relevance:** How relevant is the study to the research question? (High/Medium/Low)
Additional Notes	• **Comments:** Noteworthy comments or observations. • **Figures:** Relevant figures and tables. • **Quotes:** Relevant quotes from the study that may be useful.

details how the study defines and captures provenance information. Understanding these elements is critical for evaluating the study's approach to provenance, and provides insight into the research questions RQ1, RQ2 and RQ5. In the next section of *Provenance Utilization* the practical applications of the captured provenance data are explored. It also highlights the limitations of capturing or utilizing provenance information, which together enables us to address RQ3. The *Provenance Visualization* section examines how provenance information is presented and visualized. It describes the visualization techniques used and the tools or methods employed, which is used to answer RQ4 and RQ5. *Gaps and Future Work* identifies gaps or missing elements in the current research and suggests directions for future work. It helps highlight areas that need further investigation and helps to address RQ6. The *Quality Assessment* section evaluates the methodological rigor and relevance of the study. These evaluations ensure the inclusion of only high-quality and pertinent studies, and indicates the relevance of the study to the research questions. The final section *Additional Notes* allows for the inclusion of any other relevant information or observations. It may contain noteworthy comments, relevant figures and tables, and useful quotes from the study, providing additional context and insights. We also decided to include a comment field for each question on the form for an easier overview of the context or special remarks.

The extraction form shown in Table 7 is the final version used for the extraction, as during the review some minor modifications to the data extraction form were performed. For example the following form questions were added during the review process: Relevance, Adaptation and Figures.

The "Figure" field was added as we realized that some figures were very useful to provide context to some sections, or they simply contained helpful information. We also decided to add a cell indicating if it was connected to adaptation - either by provenance of adaptation or adaptive workflows. This to get a better overview over the actual literature on adaptation that was proving to be a smaller portion of the total amount of studies reviewed. The "Relevance" field was also used to indicate the amount of relevant information provided within the study. There was also a modification made to the "Type of study" question. Before the review process there were no specific categories planned for this field. However, after finishing the review while working on the report, we realized that it would be wise to divide the studies into a specific set of categories: General studies, Empirical studies, Conceptual studies and Adaptive-related studies.

While using the extraction initially, we answered the questions on the extraction form in great detail. However, we soon realized that it would be more efficient by providing brief answers and including additional details in the comment sections. This approach would also simplify the process of reviewing and comparing responses later on.

When reviewing studies in a systematic review it is also important to include checking techniques, to validate that the review is indeed consistently performed and and without bias. As this review process is performed by only one person this checking technique has been executed as a test-retest process, where the reviewer has performed a second extraction from a random selection of studies already reviewed. After the second extraction the results are compared to check the data extraction consistency. This second extraction was performed on one of the earlier reviewed studies by the end of the review process.

3 Results

The selected studies for the review consisted mainly of papers focused on scientific workflows, as only 5 out of the total 22 reviewed papers covered business workflows or processes. However as both workflows are similar in structures, most information are applicable in both fields. This also means that the same tools and methods can generally be used for both scientific- and business workflows. Because of this there will not be a separation of these studies based on their applied field.

The primary studies were of many different types, so to get a better understanding of how they differed and which insight we can gather from them we are going to divide them into different categories. Note that some of these primary studies can belong to more than one category.

The first category, which is also the most common type, is the *Conceptual studies*. These studies build upon previous research, using it as a foundation to propose new concepts or ideas. These concepts have not actually been fully implemented or tested yet, at the time of this study's publication.

The next category is *Empirical studies*. This study type analyses an already implemented tool or method that has been tested and analysed.

As there was a limited amount of papers related to provenance of adaptation, this was not selected as a requirement and other studies not directly related to adaptation were also included in the review. As such, we introduce a category called *Adaptation-related studies*, to indicate studies containing discussion about provenance of adaptations or adaptive workflows.

The last category is called *General studies*, and this includes studies on the general topic of provenance or workflows, as well as review papers and surveys. The contributing studies for each category can be seen in Table 8.

Table 8: Study types

Type	**References**
Conceptual	[19, 27, 4, 31, 8, 33, 32, 10, 16, 7, 25, 9, 28, 34]
Empirical	[13, 6, 39]
Adaptation-related	[19, 27, 4, 31, 8, 33, 32, 17]
General studies	[3, 11, 23, 17, 30]

Within the studies in which provenance of adaptation were discussed, the adaptation aspect was sometimes the main focus of the paper but in others it may only have been discussed briefly. There was also different terms used to describe provenance of adaptation, and the focus of adaptation would also differ. The previously mentioned form of evolution provenance was used in [17]. The term "provenance of change" was used in [31, 33] and "provenance of adaptation" was used in [33], both used to describe provenance of adaptations or changes in workflows. As stated in [33] this is necessary to enable provenance of adaptive workflows. In [4] the term "subjunctive provenance" was used to describe not the

adaptation itself, but other possible changes or events that could have happened but did not occur. In the papers [19, 27, 8, 32] there were no specific term used to describe provenance of adaptation.

The studies not directly related to adaptation were also included, if they fulfilled the selection criteria and the quality assessment. These papers were used to gain further insight into the topic of provenance in scientific and business workflows, as well as to answer the respective research questions. The sub-questions in which these papers were relevant are: RQ2, RQ3, RQ4 and RQ5 (cf. Table 1). As such, these papers will mainly provide insight into which provenance information is currently being captured, how it is being used and visualized, and what tools are being used for this purpose.

3.1 Capturing Provenance Information

When it comes to the capturing of provenance, there are two different types of studies encountered during the review that are deemed most relevant: *Conceptual studies* and *Empirical studies*. The *Conceptual studies* tend to describe the more recent works in this area, and the *Empirical studies* discuss the more established works. The *General studies* can also be used to gain insight into the general state of provenance capturing. As such, for this section there is a division of these categories to describe the capturing of provenance.

We are now going to answer RQ2: What information should be captured, and what is currently being captured? (cf. Table 1). We will first look at the information currently being captured. We will then proceed to explore the topic of what information *should* be captured, and for this we are mainly going to focus on the *Adaptation-related studies*.

3.1.1 Information currently being captured

To answer this question we are mainly looking at the *Empirical studies*, and the *General studies* will also be used. The information retrieved from these studies may not be overly specific, as the relevant information to capture will vary depending on the context and the specific workflow in question. However we have gathered from the reviewed papers that there are specific provenance types and forms that are more commonly collected.

Since this review focuses on provenance in scientific and business workflows, in the studies the mentioned provenance form collected is workflow provenance. In [17], workflow provenance is defined as meta-data collected for a workflow process that can be derived from the input, output, the model and the parameters of the workflow process. If we look more into detail at the workflow provenance captured we observe that both granularities are reported to be captured, depending on the use cases of the provenance. There are also mainly two forms that are mentioned to currently being captured, namely prospective and retrospective provenance. The only mentions of evolution provenance being captured has been in [17] which states that Kepler [20] and VisTrails [37] provides support for evolution provenance.

3.1.2 Information to be captured

To answer the question of what information should be captured (to enable the capture of adaptation), we have to look into the papers of the category *Adaptation-related studies*. It is inevitable that provenance is needed, but we will look into what exactly differs between provenance and the provenance of adaptation. In [19] it is mentioned that capturing the provenance of the adaptation steps made within the workflow is necessary, and to further explore this we will examine exactly what information this entails.

Some studies identify that there are new forms of provenance needed to be collected to enable adaptation, additionally to the forms defined by [17]. As before mentioned there are three provenance forms identified: prospective, retrospective and evolution provenance. In [4] a fourth form of provenance is also identified, namely subjunctive provenance. This form of provenance describes what could happen during the implementation of a process. The authors provide an effective demonstration of the various provenance forms using the process of building an IKEA table as an example.

Using this example, subjunctive provenance could be explained by the fact that some tool described in the IKEA table manual is not included in the kit, which may lead to the customer using another tool or method to assemble the different parts of the table. It could also be used the other way around, if such a modification has taken place subjunctive provenance may be used to describe what could have happened would this change not have occurred. Subjunctive provenance is thereby needed to identify potential branches that could emerge within a workflow, or to look back at a process and see what could have happened if other choices were made.

In [33] there has also been an additional form of provenance identified, namely ad-hoc provenance. Provenance of change has been subdivided in [33] into the forms evolution provenance (originally mentioned in [17]) and ad-hoc provenance. The difference between these two are identified as evolution provenance describe adaptations to the workflow model itself, and ad-hoc provenance describes the adaptations in a workflow instance. Using the same example as in [4] with the IKEA table, evolution provenance can be compared to the company replacing the provided screws in the table package for nails. This would enable customers to use a hammer for assembly instead of a screwdriver, generating a change in the building process for the table, affecting all customers. Ad-hoc provenance on the other hand could represent one specific customer not finding the right sized screwdriver, and therefore opting to exchange the screws for nails that they already possessed. As such, this does not affect the general building process for the table, only for this specific instance of it.

3.2 Usage and Visualization

There are many use cases for provenance mentioned in the studies, but the most commonly discussed utilization is for the purpose of reproducibility within workflows. Within the scientific community reproducible research is fundamental and within scientific and business workflows provenance information is the very basis of this important feature.

Within the area of collaborative adaptive workflows, provenance of adaptation is especially required to ensure reproducibility according to [32, 33]. In these papers the authors assert that in order to reproduce data processing pipelines between collaborating organisations it is necessary to record a description of the change being performed, as well as the new and the old version of the workflow.

Except for reproducibility there are multiple other use cases for provenance. Figure 3 presents the use cases discussed in the studies, along with the corresponding studies where these use cases were mentioned. Many use cases stated in Figure 3 are related to each other. For example: analyzing and validating the data and its derivation can help researchers explain unexpected results or identifying and handling errors. This makes it possible to further prevent fraud and ensure trust within the collaborators, as all the results are traceable. Moreover, handling errors and preventing fraud helps maintain quality throughout the experiments thereby facilitating quality control. As such, collecting and documenting provenance enables a lot of use cases with different advantages for researchers.

When it comes to visualization of provenance there are different approaches. The majority of primary studies discussing visualization mention some form of graph-representation to visualize the provenance. The use of a Directed Acyclic Graph (DAG) is described in [28, 6, 34, 13], although these graphs are constructed in different ways. For example the visualization graphs in [6] are made of graphs where each node represents data, collection and parameter items produced by (or provided to) the workflow run. The edges maps each node to the set of nodes and events involved in its creation, showing how one item is derived from others or influenced by specific events.

The graph described in [13] are described as a rooted tree, where each node represents a version of the workflow. The edges between these nodes then describe the action needed for the derivation of one version from another. Some tools and methods discussed in the studies provide support for PROV, which is a provenance standard that can implicitly be visualised as a graph.

Some less common visualization techniques mentioned are the use of provenance reports [16], and a database-visualization [6].

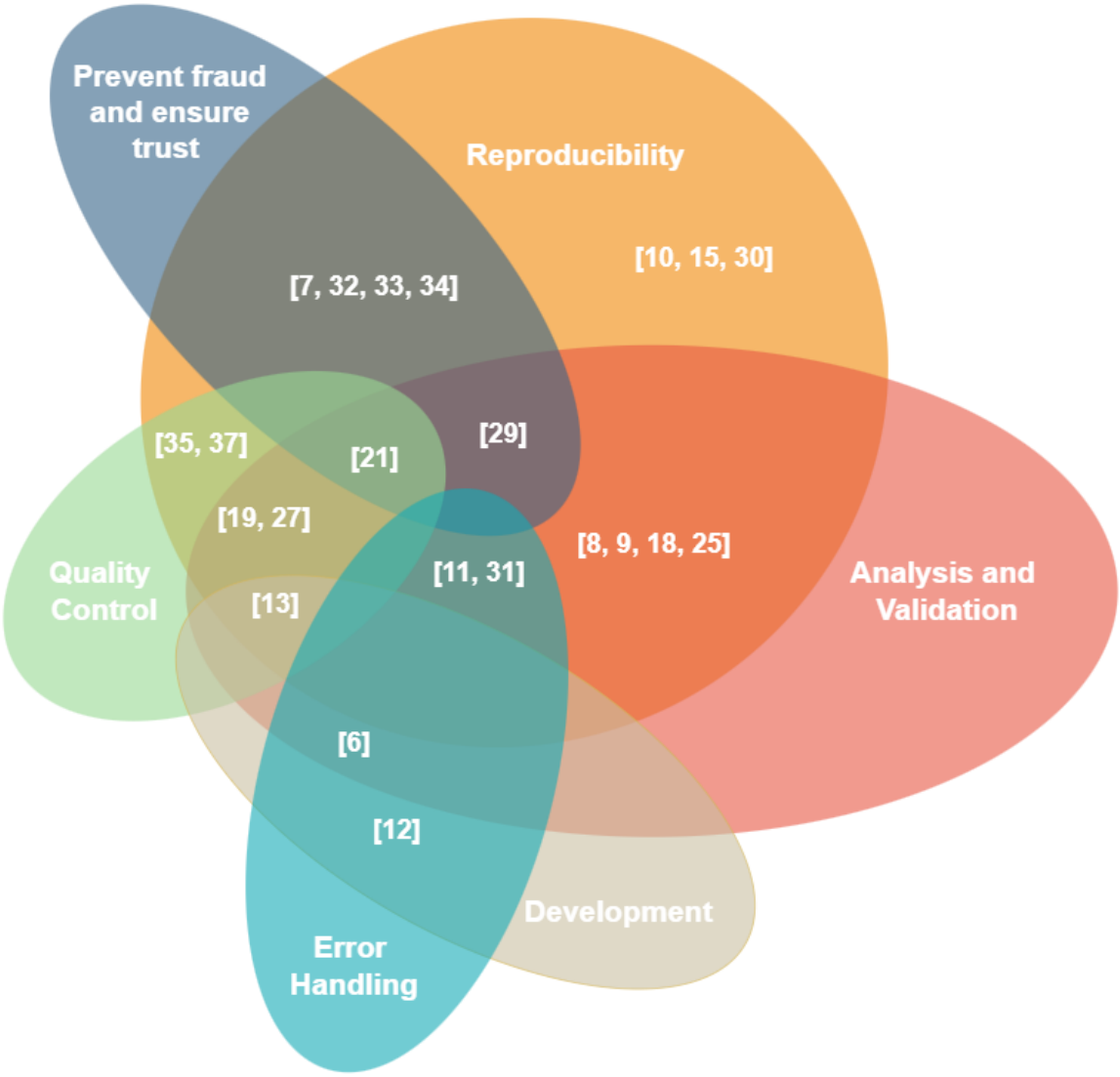

Figure 3: Use cases in the literature

3.3 Tools and works

There has been many different tools and methods encountered during the review, and once again it is necessary to divide them into two categories: implemented tools and methods, and conceptual tools and methods. The implemented tools are tools and methods that are fully implemented and officially released. Conceptual tools are defined as those that are either not fully implemented or consist of elements that are still only theoretical.

We will begin looking at the implemented tools and methods as these will give us a general view of what is accessible right now. Later we will look into conceptual tools and methods which will be more related to the capturing of provenance of adaptations. The tools discussed in this section may be discussed in great detail in the studies or simply be briefly mentioned. If briefly mentioned, additional research has been made to answer the relevant questions such as: What is the main purpose of this tool? Which provenance standards are supported? Etc.

3.3.1 Implemented tools and methods

During the review, various provenance tools designed for different purposes were encountered. A collection of the reviewed implemented tools and methods is illustrated in Table 9. The tools included in these tables are the tools mentioned in the studies that contain some support for provenance. It is stated what the main purpose of the tool is in regards to provenance, and whether there is support for some provenance standard. Under "Type," it is specified whether the tool is a WfMS, incorporated with a WfMS, or a standalone tool. "Last Change" indicates the most recent year when the code in the open repository was updated. Lastly it is stated in which papers this tool or method was mentioned.

Table 9: Implemented tools. C: Capture, M: Manage, V: Visualize, E: Edit

Name	Purpose	Standard	Type	Last Change	References
Kepler	C	PROV, OPM	WfMS	2021	[33, 28, 6, 7, 11, 3, 17]
Prov Viewer [1]	V	PROV	Standalone	2020	[33, 39]
ProvViz [1]	V, E	PROV, RDF	Standalone	2021	[33, 39]
ReproZip	C, M	None	Standalone	2024	[8]
Taverna	C, M	PROV, OPM	WfMS	2020	[7, 11, 3, 32, 17]
VisTrails	C, M, V	OPM	WfMS	2017	[8, 34, 13, 7, 11, 17, 30]

As can be seen in Table 9, most of these tools have been developed for the main purpose of managing provenance, collection of provenance or for visualization. The most mentioned tools are: Kepler [20], VisTrails [37] and Taverna [35]. Of these tools, only Kepler is still being maintained as both VisTrails and Taverna are no longer maintained according to their respective websites or specifications. According to [17], evolution provenance is supported by Vistrails and Kepler, but none of the tools are reported so support ad-hoc provenance. Moreover, all the implemented tools mentioned in this table were found to be open source.

3.3.2 Conceptual tools and methods

As previously stated, many of the primary studies used in this review are conceptual studies, introducing a conceptual or theoretical idea that has not yet been fully implemented. These concepts are important to distinguish from the implemented tools and methods, as these have not yet been thoroughly tested. These concepts do however contain potential and may provide methods that could enable the capture and management of provenance of adaptation, which is something that the already implemented tools and methods have not been able to

fully satisfy. Therefore this category is very important to be able to answer the main research question. An overview of the conceptual tools and methods encountered during this review are presented in Table 10. The columns are the same as Table 9, with the addition of the "Adaptation" and "Open Source" columns, indicating whether or not there is expressed support for adaptations, and if the tool is open source. Instead of noting the last year of change, we also consider the "Last Work". Since not all of these tools are implemented, this can also indicate when the most recent paper discussing the tool was published.

Table 10: Conceptual tools. C: Capture, M: Manage, V: Visualize, S: Storage

Name	Purpose	Adaptation	Standard	Type	Open source	Last Work	References
AVOCADO	V	×	None	Standalone	×	2017	[34, 39]
PRISM	C, M, V, S	✓	None	Standalone	✓	2023	[27]
Provenance Holder	C	✓	PROV	Incorporated	✓	2024	[32, 33, 31]
ProvSearch	M, S	×	None	Incoprorated	×	2014	[10]
SAMbA-RaP	C, S	×	PROV	Incorporated	✓	2020	[16]
Secure scientific workflow data provenance framework	C, S	×	ProVOC, RDF	Standalone	×	2023	[23]

From Table 10 we can derive that only two tools currently indicate some form of support for provenance of adaptation: PRISM and Provenance Holder. PRISM introduces a design enabling the storing of provenance on a decentralized ledger in the form of a blockchain. It is meant to provide a flexible framework to support the dynamic nature of scientific workflows. To enable the support for workflow modifications, PRISM uses Invalidation and Modification Blocks. A scientist needs to get the agreement of at least 50% of the scientist to be able to submit an invalidation transaction, or a modification transaction. After this a Data Invalidation or Modification Block will be added to the ledger. When the Invalidation Block gets propagated into the network, the relevant data will be flagged as invalid. Following this invalidation, the workflow task which produced the invalid data will be recomputed. As a consequence of this, every other workflow task relying on the invalid data must also be invalidated and recomputed. If there is a need for addition or subtraction of workflow tasks, a Modification Block is to be used. This provides scientists the possibility to adapt workflows during the experiments. The use of blockchain in this implementation simultaneously ensures comprehensive and transparent provenance records [27].

According to [33] the Provenance Holder will have support for the coarse-grained granularity, and the prospective and evolution provenance forms. More specifically it will support both workflow evolution provenance and provenance of ad-hoc workflow change, according to the previously introduced definitions. It is also stated that support for retrospective and fine-grained provenance could also be enabled with detailed workflow execution traces. The Provenance Holder collects provenance on a very detailed level during the execution and adaptation of workflows. The data collected is highly specific and includes information about each individual activity within the workflows. The specific provenance information for a workflow adaptation might consist of a description of the specific modification, the new version of the workflow, as well as a reference to the old (preceding) version. The adaptations to the workflow model are stored in an object type called provenance information object for adaptation. When an instance migration occurs, the entire workflow model is captured. However, if ad-hoc workflow changes happen, only the specific modification is captured. Before storing the provenance the data gets validated. Similarly to PRISM the

Provenance Holder suggests the use of an immutable public ledger, for example blockchain, but instead of storing the provenance it is used for time-stamping. The primary use for this time-stamping is to attribute to the fact that something was known for example before a certain action.

In [8] they also describe a method of descriptor-space containing all the necessary parameters for the workflow within descriptors. For each descriptor there is a name, a value and a decay-parameter. This decay-parameter is tracking whether the parameters are accessible, as well as if there are changes within the parameters. For the workflow to be fully documented and reproducible all parameters need to be known or stored. Therefore this descriptor-space method makes it possible to analyse whether or not a workflow is fully reproducible and if there has been any changes made to the parameters. One example of this could be if a workflow is depending on data subtracted from a database that is continuously changing, the decay-parameter will indicate that a change has been made and this adaptation can be documented.

There is also a new scientific workflow data provenance framework proposed in [23], which is based on the provenance model ProVOC and blockchain. Using blockchain, it links the provenance information to verify its reliability and ensure that it has not been altered or tampered with.

3.3.3 Provenance standards

There are many different provenance representation models or standards used to represent provenance. Initially, Resource Description Framework (RDF) [15] was used to present provenance. RDF is a data model developed by World Wide Web Consortium (W3C) for meta-data and a key element of the semantic web. RDF represents data as triples, each consisting of a subject, predicate, and object.

Following this, one of the first models developed specifically for provenance in 2007 was the Open Provenance Model (OPM) [29], which is a Graph-Based Model. OPM did however come with some problems and as a result another more refined and detailed standard was proposed by W3C, called PROV [38]. The very core of the PROV standard is the data model The PROV Data Model (PROV-DM). PROV-DM can also be mapped to RDF via PROV-O.

Another notable model is Open Provenance Model for Workflows (OPMW), which serves as an extension of the OPM standard specifically designed to describe the provenance of scientific workflows. Also highlighted in the studies is ProvOne [5], which also represents scientific workflow provenance. Notably, ProvOne is compatible with PROV-DM, ensuring interoperability and flexibility in how provenance information is captured and utilized across different systems. Together, these models provide robust frameworks for managing the complexities of scientific workflows.

According to [23] PROV-DM, OPMW and ProvOne all have the ability to capture, store, and search the provenance of a workflow, as well as trace it in a standard, machine-readable format. ProvOne is also supporting three forms of provenance: prospective, retrospective and evolution [17].

4 Discussion

In this literature review report we have examined the state of the art of provenance in (adaptive) scientific and business workflows. We have explored the capture of provenance and discovered that there are a lot of different tools and methods available. However these methods are generally only capturing prospective and/or retrospective provenance and there is a lack of support for provenance of adaptations. We have also looked into how provenance is visualized. There are a lot of different tools and approaches available with the most common approach being a graph representation of some sort, for example a DAG.

Provenance has many areas of use, the most mentioned being reproducibility. Other use-cases identified are: Prevent fraud and ensure trust, Analysis and validation, Quality control, Error handling and Development. While many of these areas are interconnected, they can also be viewed independently.

We have examined different tools mentioned in the reviewed papers, and divided them into two categories: implemented- and conceptual tools. Within the implemented tools there is some support for evolution provenance, more specifically by Vistrails and Kepler. We can also argue that some of the visualizing tools with support for a provenance standard can support evolution provenance. This is due to the fact that some of these provenance standards are able to model provenance including the adaptations. The conceptual tools on the other hand also introduce some support for provenance of adaptations (see PRISM [27] and the Provenance Holder [33]). There are also tools for complex provenance visualization (AVOCADO [34]), as well as other managing and storing tools.

Using the information provided we are able to answer the main research question: How do we capture provenance information of adaptive scientific and business workflows?

We have examined two different forms of adaptive provenance: evolution and ad-hoc. To enable the capture of provenance information of adaptive scientific- and business workflows, it is important that we provide support for the capture of both evolution and ad-hoc provenance. As the most common use-case of provenance is reproducibility, it is crucial to enable the capture and managing of all forms of provenance, including provenance of adaptation. Without this we can not ensure full reproducibility for adaptive workflows. Some implemented tools (Kepler and Vistrails) do provide support for the capture of evolution provenance, but no support of ad-hoc provenance has been reported. There are also tools under development with support for provenance of adaptations, such as PRISM [27] and the Provenance Holder [33], although they are not fully implemented yet. The Provenance Holder reports support of both evolution and ad-hoc provenance, while PRISM does not define the supported provenance forms.

When it comes to the visualization of provenance of adaptations, there are tools available that supports the PROV standard (ProvViz [39] and Prov Viewer [36]). Since this standard is flexible enough to capture the provenance and the adaptations, these tools are already compatible with provenance of adaptation, provided it is represented in a PROV format.

4.1 Threats to validity

One of the primary threats to the validity of this literature review is the relatively small number of papers available on the topic of provenance of adaptation. This limitation can impact the comprehensiveness of our findings as well as potential biases in the results. As provenance of adaptation is a relatively new and evolving area of research, the literature is still developing and some studies might be in progress or unpublished.

4.2 Future work

There is still a lot of open questions and future work available within this topic. All the areas mentioned or proposed in the conceptual studies or the conceptual tools which are not fully implemented yet serves as a base for future work. The threats to validity mentioned is also an argument for future work, as there is an apparent need for further investigations in this area.

There is currently a somewhat lack of tools or methods provided to enable the capture of evolution provenance. Moreover there have been additional forms of provenance identified: ad-hoc provenance [33] and subjunctive provenance [4]. The ad-hoc provenance has been a subject of the work by L. Stage and D. Karastoyanova, who are also developing a provenance capture tool that will provide support for this form of provenance. Subjunctive provenance could be further explored in relation to evolution and ad-hoc provenance, as this topic is not covered in the works by Bettivia et al. There are also possibilities to further examine how subjunctive provenance could be captured and visualized.

As more forms of provenance are identified, the volume of provenance information that needs to be captured, stored, and visualized also increases. This drives the development of new tools and methods capable of managing these complex data systems. Another key area for future work is determining which specific provenance information is necessary to ensure reproducibility or to meet the particular needs of the intended use case.

5 Acknowledgements

I would like to extend my heartfelt gratitude to my first supervisor, Prof. Dimka Karastoyanova, for her encouragement and for giving me the opportunity to work on such an intriguing topic. I also wish to express my sincere appreciation for my second supervisor Ludwig Stage for his invaluable assistance throughout the entire process, offering both his time and support. This guidance has been incredibly valuable and without it, this work would not have reached its current state.

References

[1] Scopus. https://www.scopus.com/. Accessed 2024-08-05.

[2] Web of science. www.webofscience.com. Accessed 2024-08-05.

[3] Malcolm Atkinson, Sandra Gesing, Johan Montagnat, and Ian Taylor. Scientific workflows: Past, present and future. *Future Generation Computer Systems*, 75:216–227, Oct 2017.

[4] Rhiannon Bettivia, Yi-Yun Cheng, and Michael Gryk. What does provenance lack: How retrospective and prospective met the subjunctive. *Lecture notes in computer science*, page 74–82, Jan 2023.

[5] Rhiannon Bettivia, Yi-Yun Cheng, and Michael Robert Gryk. Provone. *Documenting the Future: Navigating Provenance Metadata Standards. Synthesis Lectures on Information Concepts, Retrieval, and Services*, page 41–56, Jan 2022.

[6] Shawn Bowers, Timothy M. McPhillips, and Bertram Ludäscher. Provenance in collection-oriented scientific workflows. *Concurrency and Computation: Practice and Experience*, 20(5):519–529, 2008.

[7] Anila Sahar Butt and Peter Fitch. A provenance model for control-flow driven scientific workflows. *Data & Knowledge Engineering*, 131-132:101877, Jan 2021.

[8] Anna Bánáti, Péter Kacsuk, and Miklós Kozlovszky. Reproducibility analysis of scientific workflows. *Acta Polytechnica Hungarica*, 14(2), 2017.

[9] A Chebotko, S Lu, S Chang, F Fotouhi, and P Yang. Secure abstraction views for scientific workflow provenance querying. *IEEE Transactions on Services Computing*, 3(4):322–337, Oct 2010.

[10] Flavio Costa, Daniel de Oliveira, and Marta Mattoso. Towards an adaptive and distributed architecture for managing workflow provenance data. *2014 IEEE 10th International Conference on e-Science*, Oct 2014.

[11] Susan B Davidson and Juliana Freire. Provenance and scientific workflows: challenges and opportunities. *International Conference on Management of Data*, Jun 2008.

[12] DBLP. dblp: computer science bibliography. https://dblp.org/. Accessed 2024-07-24.

[13] Juliana Freire, Cláudio T Silva, Steven P Callahan, Emanuele Santos, Carlos Scheidegger, and Huy T Vo. Managing rapidly-evolving scientific workflows. *Lecture Notes in Computer Science*, vol 4145:10–18, Jan 2006.

[14] Diimitrios Georgakopoulos, Mark Hornick, and Amit Sheth. An overview of workflow management: From process modeling to workflow automation infrastructure. *Distributed and Parallel Databases*, 3(2):119–153, Apr 1995.

[15] RDF Working Group. Rdf - semantic web standards. https://www.w3.org/RDF/, 2014. Accessed 2024-08-03.

[16] Thaylon Guedes, Lucas Amaral Martins, Maria Luiza Falci, Vitor Silva, Kary, Marta Mattoso, Marcos, and Daniel De Oliveira. Capturing and analyzing provenance from spark-based scientific workflows with samba-rap. *Future Generation Computer Systems*, 112:658–669, Nov 2020.

[17] Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. A survey on provenance: What for? what form? what from? *The VLDB Journal*, 26(6):881–906, Oct 2017.

[18] IEEE. Ieee xplore digital library. https://ieeexplore.ieee.org/Xplore/home.jsp. Accessed 2024-07-24.

[19] Dimka Karastoyanova and Ludwig Stage. Towards collaborative and reproducible scientific experiments on blockchain. *Lecture notes in business information processing*, 316:144–149, Jan 2018.

[20] Kepler. The kepler project — kepler. https://kepler-project.org/. Accessed 2024-08-04.

[21] Barbara Ann Kitchenham and Stuart Charters. Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report, 07 2007.

[22] Frank Leymann and Dieter Roller. *Production Workflow: Concepts and Techniques*. Prentice Hall PTR, 2000.

[23] Songhai Lin, Hong Xiao, Wenchao Jiang, Dafeng Li, Jiaben Liang, and Zelin Li. A survey of provenance in scientific workflow. *Journal of High Speed Networks*, 29(2):129–145, Apr 2023.

[24] Bertram Ludäscher, Shawn Bowers, and Timothy McPhillips. Scientific workflows. *Encyclopedia of Database Systems*, page 2507–2511, 2009.

[25] Anderson Marinho, Leonardo Murta, Cláudia Werner, Vanessa Braganholo, Sérgio Manuel Serra da Cruz, Eduardo Ogasawara, and Marta Mattoso. Provmanager: a provenance management system for scientific workflows. *Concurrency and Computation: Practice and Experience*, 24(13):1513–1530, Oct 2011.

[26] J. P. Mesirov. Accessible reproducible research. *Science*, 327(5964):415–416, Jan 2010.

[27] Matthew Miller, Skarlet Williams, Gaby G Dagher, and Min Long. Prism: A blockchain-enabled reputation-based consensus for enhancing scientific workflow provenance. *IEEE Computer Society*, Nov 2023.

[28] Sonia Mitchell, Andrew Lahiff, Nathan Cummings, J Hollocombe, Bram Boskamp, Ryan Field, Dennis Reddyhoff, Kristian Zarebski, Antony Wilson, B Viola, Martin Burke, Blair Archibald, Paul R Bessell, Richard E Blackwell, Lisa Boden, Alys Brett,

Sam Brett, Ruth Dundas, Jessica Enright, and Alejandra Gonzalez-Beltran. Fair data pipeline: provenance-driven data management for traceable scientific workflows. *Philosophical Transactions of the Royal Society A*, 380(2233), Aug 2022.

[29] OPM. The opm provenance model (opm). https://openprovenance.org/opm/. Accessed 2024-08-03.

[30] Eric D. Ragan, Alex Endert, Jibonananda Sanyal, and Jian Chen. Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):31–40, Jan 2016.

[31] Ludwig Stage. Trusted provenance of collaborative, adaptive, process-based data processing pipelines. *Lecture notes in business information processing*, 498:363–370, Jan 2024.

[32] Ludwig Stage and Dimka Karastoyanova. Provenance holder: Bringing provenance, reproducibility and trust to flexible scientific workflows and choreographies. *Lecture notes in business information processing*, 362:664–675, Jan 2020.

[33] Ludwig Stage and Dimka Karastoyanova. Trusted provenance of automated, collaborative and adaptive data processing pipelines. *arXiv*, Oct 2023.

[34] H. Stitz, S. Luger, M. Streit, and N. Gehlenborg. Avocado: Visualization of workflow-derived data provenance for reproducible biomedical research. *Computer Graphics Forum*, 35(3):481–490, Jun 2016.

[35] Taverna. https://www.w3.org/2011/prov/wiki/TavernaProvenance, 2024. Accessed 2024-07-24.

[36] Prov Viewer. https://gems-uff.github.io/prov-viewer/, 2018. Accessed 2024-08-03.

[37] VisTrails. Vistrails — main page. https://www.vistrails.org//index.php/Main_Page. Accessed 2024-08-04.

[38] W3C. Prov-overview. https://www.w3.org/TR/prov-overview/. Accessed 2024-08-03.

[39] Ben Werner and Luc Moreau. Provviz: An intuitive prov editor and visualiser. *Lecture notes in computer science*, 12839:231–236, Jan 2021.

[40] Mathias Weske. *Business Process Management - Concepts, Languages, Architectures, Third Edition*. Springer, 2019.

[41] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, and Alejandra Gonzalez-Beltran. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), Mar 2016.