REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

DATA PROVENANCE VISUALIZATION METHODOLOGIES

İlkay Melek YAZICI

DOCTOR OF PHILOSOPHY THESIS
Department of Computer Engineering
Program of Computer Engineering

Supervisor
Assoc. Prof. Dr. Mehmet S. AKTAŞ

January, 2023

REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

DATA PROVENANCE VISUALIZATION METHODOLOGIES

A thesis submitted by İlkay Melek YAZICI in partial fulfillment of the requirements for the degree of **DOCTOR OF PHILOSOPHY** is approved by the committee on in Department of Computer Engineering, Program of Computer Engineering.

Assoc. Prof. Dr. Mehmet S. AKTAŞ
Yildiz Technical University
Supervisor

Approved By the Examining Committee

Assoc. Prof. Dr. Mehmet S. AKTAŞ, Supervisor
Yildiz Technical University _____

Assoc. Prof. Dr. Mehmet GÖKTÜRK, Member
Gebze Technical University _____

Dr. Yunus Emre SELÇUK, Member
Yildiz Technical University _____

Prof.Dr. Oya KALIPSIZ, Member
Yildiz Technical University _____

Assoc. Prof. Dr. Hasan SÖZER, Member
Ozyegin University _____

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of Data Provenance Visualization Methodologies supervised by my supervisor, Assoc. Prof. Dr. Mehmet S. AKTAŞ. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

İlkay Melek YAZICI

Signature

Dedicated to my dear Mom and Dad;

ACKNOWLEDGEMENTS

Words cannot express my gratitude to my supervisor, Assoc. Prof. Dr. Mehmet AKTAŞ for his invaluable patience and endless guidance. You made this work possible. Thank you for your invaluable supervision, support and tutelage during the course of my PhD degree.

I wish to thank my committee members who were more than generous with their expertise and precious time. A special thanks to Dr. Yunus Emre SELÇUK and Assoc. Prof. Dr. Mehmet GÖKTÜRK for their encouragement, and most of all patience throughout the entire process. I also could not have undertaken this journey without my defense committee, who generously provided knowledge and expertise.

Additionally, I'm grateful to my classmates and cohort members, especially Erkan KARABULUT, for his crucial assistance at the beginning of this thesis project. Also deserving of praise are the university study participants, research assistants, and office colleagues who helped and motivated me along the way.

Lastly, I would be remiss in not mentioning my family, especially my parents, Nazende-Murat Yazıcı, my sister Esra, my brother Emrah, and my little love Ozan Emre. Their belief in me has kept my spirits and motivation high during this process. And my first teacher Melahat POLAT who made me love research. My friends, whom I neglected during my busy times, are grateful to you for the fun and emotional support you have provided along the way.

İlkay Melek YAZICI

TABLE OF CONTENTS

LIST OF FIGURES viii

LIST OF TABLES x

ABSTRACT xi

ÖZET xii

1 INTRODUCTION 1

1.1 Literature Review . . . 3
1.2 Objective Of The Thesis . . . 5
1.3 Hypothesis . . . 7
1.4 Statement of Research Problem . . . 7
1.5 Thesis Scope . . . 8
1.5.1 Provenance Visualization Requirements . . . 8
1.5.2 Application Use Domains . . . 10
1.5.3 Thesis Contributions . . . 14
1.6 Organization of Thesis . . . 15

2 FUNDAMENTAL CONCEPTS 16

2.1 Data Provenance . . . 16
2.1.1 Provenance Taxonomy . . . 18
2.2 Provenance Ontology Model . . . 19
2.2.1 The World Wide Web Consortium (W3C) Working Group . . . 20
2.2.2 W3C PROV Family . . . 21
2.2.3 W3C PROV-O Ontology Model . . . 21
2.3 Data Provenance Visualization . . . 24
2.4 Systematic Literature Review (SLR) . . . 24
2.5 HCI Design . . . 25
2.6 Usability Testing . . . 26
2.6.1 System Usability Scale . . . 27
2.6.2 SU Scale Evaluation: . . . 27

3 SYSTEMATIC LITERATURE REVIEW **29**
3.1 Overall Approach . . . 29
3.2 Systematic Literature Review on Data Provenance Visualization . . . 29
3.2.1 Research Questions . . . 30
3.2.2 Search String and Data Sources . . . 30
3.2.3 Study Selection and Data Extraction . . . 31
3.2.4 Data Extraction Results . . . 32
3.2.5 Discussion of Research Questions . . . 33
3.2.6 Threats to Validity and Limitations . . . 39
3.3 Related Work and Our Contributions . . . 39

4 DATA PROVENANCE VISUALIZATION METHODOLOGIES **41**
4.1 Overall Approach . . . 41
4.1.1 Research Questions . . . 41
4.1.2 Requirements and Use Cases . . . 42
4.2 Data Provenance Visualization Methodologies . . . 46
4.2.1 Proposed Visualization Methodologies . . . 47
4.2.2 Datasets . . . 62
4.2.3 Prototype Implementation . . . 64
4.2.4 Experimental Study Results . . . 67
4.2.5 Discussion on Research Questions . . . 73
4.2.6 Threats to The Validity of the Research Results . . . 74
4.3 Related Work and Our Contributions . . . 75

5 USABILITY STUDY ON PROVENANCE DATA VISUALIZATION TOOLS **80**
5.1 Overall Approach . . . 80
5.1.1 Research Questions . . . 81
5.2 Usability Study . . . 81
5.2.1 Methodology . . . 81
5.2.2 Usability Testing Sessions . . . 82
5.2.3 Usability Test Participants . . . 83
5.2.4 User Surveys and Evaluation . . . 83
5.2.5 Experimental Study Results . . . 85
5.2.6 Discussion on Research Questions . . . 91
5.2.7 Threats to The Validity of the Research Results . . . 94
5.3 Related Work and Contributions . . . 94

6 RESULTS AND DISCUSSION **96**
6.0.1 Future Research Opportunities . . . 100

REFERENCES **102**

A USABILITY STUDY **109**

A.1 Functional Task Level Usability Study 109

A.1.1 Backward Provenance Dependency Querying 109

A.1.2 Streaming Provenance Visualization 110

A.1.3 Provenance Data Summarization Visualization 111

A.1.4 Provenance Data Comparison Visualization 112

PUBLICATIONS FROM THE THESIS **115**

LIST OF FIGURES

Figure 1.1 Sample Process-Artifact Relations in time 8
Figure 1.2 Sample Individual Process-Artifact Relations in time 9
Figure 1.3 Sample Lineage Process order 9
Figure 1.4 Conceptual (“napkin drawing”) view of the Promoter Identification Workflow (PIW) [44] 11
Figure 1.5 PIW implemented in Kepler [44] 11
Figure 1.6 An example social media provenance data[29] 13
Figure 2.1 Data lineage illustration 17
Figure 2.2 Provenance Taxonomy 18
Figure 2.3 W3C PROV Family 22
Figure 2.4 W3C PROV-O components 23
Figure 2.5 SUS Assessment Question Set 28
Figure 2.6 SUS Score Categorization 28
Figure 3.1 SLR process steps 32
Figure 3.2 Publications in years 34
Figure 4.1 Use cases for interoperable ontology-based provenance visualization 42
Figure 4.2 Use cases for user-centric provenance visualization 43
Figure 4.3 Use cases for provenance data summarization 44
Figure 4.4 Use cases for provenance data comparison 45
Figure 4.5 Use cases for data provenance dependency querying 45
Figure 4.6 Use cases for streaming provenance data visualization 46
Figure 4.7 PROV-O items’ individual representation 47
Figure 4.8 PROV-O compatible data importation 48
Figure 4.9 Data conversion sample 48
Figure 4.10 Example for node property window 50
Figure 4.11 Provenance data summarization sample 51
Figure 4.12 Pseudo code for provenance summarization algorithm 53
Figure 4.13 DCA-Prov provenance data comparison 54
Figure 4.14 DCA_{label} provenance data comparison 54
Figure 4.15 Enhanced-DCA Node Attendance Calculation Pseudo code 57
Figure 4.16 Backward Dependency Query Sample 60

Figure 4.17 Backward dependency querying pseudo code 61
Figure 4.18 Provenance stream visualization solution 63
Figure 4.19 Visualization Prototype 65
Figure 4.20 Spark clustering environment 66
Figure 4.21 Provenance visualization streaming flow 67
Figure 4.22 Elbow function results for optimum cluster number 68
Figure 4.23 pySpark KMeans Clustering scalability results 69
Figure 4.24 pySpark KMeans-RDD clustering scalability results 69
Figure 4.25 pySpark Bisecting-KMeans clustering scalability results 69
Figure 4.26 pySpark Gaussian Mixture Model clustering scalability results . . . 70
Figure 4.27 DCA_{label} (up) and DCA_{props} before/after graph comparison (down) on sample NASA AMSR-E dataset 71
Figure 4.28 100k size E_BDM and C_BDM comparison 72
Figure 4.29 400k size E_BDM and C_BDM comparison 73
Figure 4.30 1M size E_BDM and C_BDM comparison 73
Figure 5.1 Usability study methodology 82
Figure 5.2 Task-1 Usability Test Responses 86
Figure 5.3 Task-2 Usability Test Responses 86
Figure 5.4 Task-3 Usability Test Responses 87
Figure 5.5 Task-4 Usability Test Responses 88
Figure 5.6 General Task Level Satisfaction Diagram 89
Figure 5.7 General Task Level User Satisfaction for Each Task 90
Figure 5.8 SUS Mean Rating(left) and Percent Agree(right) 91

LIST OF TABLES

Table 2.1 Likert Scale values 27
Table 3.1 Research Questions. 30
Table 3.2 Digital Data Sources 31
Table 3.3 Categorical Data and Reviewed Studies Reference mapping 33
Table 3.4 Journals and Conferences That Reviewed Papers Published Under . 34
Table 3.5 Study and Domain/Use Case Relation 35
Table 3.6 Provenance data model specification proposed in the studies 36
Table 3.7 Visualization feature comparison 37
Table 3.8 Interactive visualization Components comparison 38
Table 4.1 Research Questions 41
Table 4.2 Differences and similarities of DCA algorithms 55
Table 4.3 Comparison of BDM Algorithms 62
Table 4.4 Scalability Data Sets 68
Table 4.5 Clustering Algorithms and parameters for scalability study 69
Table 4.6 Graph comparison accuracy results 70
Table 4.7 Graph comparison results 71
Table 4.8 Performance data edge and node information 72
Table 5.1 Research Questions 81
Table 5.2 Usability study test participants 83
Table 5.3 Functional Level Usability Task 1 – Mean Task and Percent Agree Ratings 85
Table 5.4 Functional Level Usability Task 2 – Mean Task and Percent Agree Ratings 86
Table 5.5 Functional Level Usability Task 3 – Mean Task and Percent Agree Ratings 87
Table 5.6 Functional Level Usability Task 4 – Mean Task and Percent Agree Ratings 87
Table 5.7 General Usability – Mean Task and Percent Agree Ratings 88
Table 5.8 SUS Evaluation – Mean Task and Percent Agree Ratings 89
Table 5.9 Tasks and the durations 91

ABSTRACT

Data Provenance Visualization Methodologies

İlkay Melek YAZICI

Department of Computer Engineering
Doctor of Philosophy Thesis

Supervisor: Assoc. Prof. Dr. Mehmet S. AKTAŞ

Data provenance has created a growing need for methodologies that enable end-users to analyze the data life cycle and take appropriate action. In the era of big data, businesses' global data output grows daily. As data rises, so are the metadata on the data's source and life cycle. As a result, this calls for technologies that might improve data understanding and interpretation through data provenance. The difficulty of collecting graph-based data from scientific workflows and enabling necessary visualization techniques like graph comparison, data summarization, backward and forward dependency querying, and streaming data visualization are addressed in this paper. A visualization tool is used to implement the W3C-PROV-O provenance standard and evaluate the applicability of suggested techniques. A detailed Systematic Literature Review study was conducted intensively. The performance of the suggested methods is investigated using a sizable provenance data set. This study also goes into great depth on a thorough usability analysis of the prototype visualization tool. According to the results, the processing overhead of the suggested visualization approaches is negligible and they can be used.

Keywords: data provenance, data visualization, big provenance, big provenance visualization, human computer interaction

YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

ÖZET

Veri Provenansı Görselleştirme Metodolojileri

İlkay Melek YAZICI

Bilgisayar Mühendisliği Anabilim Dalı
Doktora Tezi

Danışman: Doç. Dr. Mehmet S. AKTAŞ

Veri provenansı son kullanıcıların veri yaşam döngülerini yönetmek ve analiz etmek için yeni teknolojilerin geliştirilmesi gereğini doğurmuştur. Veri boyutu arttıkça, veri kaynağı ve verilerin yaşam döngüsüne ilişkin meta veri boyutu da aynı oranda artar. Bu nedenle veri provenansı üzerinde analiz ve madenciliği destekleyen innovasyonlara ihtiyaç duyulmaktadır. Bu tez çalışmasında bilimsel iş süreç verisinin grafikler şeklinde çıkarılması ve grafik kıyaslama, özetleme, ileriye-geriye yönelik bağımlılık sorgulama gibi bilim adamlarının ihtiyaç duyacağı görselleştirme yaklaşımlarının araştırılması ve gelişirilmesi hedeflenmektedir. Çalışılan görselleştirme metodolojileri ve sunulan algoritmaların uygulanabilirliğini gözlemlemek ve değerlendirmek için W3C PROV provenans spesifikasyonunun kullanıldığı bir görselleştirme aracı geliştirilmiştir. Ayrıntılı bir Sistematik Literatür Taraması ile bu alanda yapılmış geçmiş çalışmaların belirlenen parametreler doğrultusunda değerlendirilmesi yapılmıştır. Sunulan yaklaşımlar geniş ölçekli bir kullanılabilirlik çalışması ile sistematik bir şekilde değerlendirilmiş ve SUS metrikleri referans alınarak kullanılabilirliğinin başarısı raporlanmıştır. Elde edilen sonuçlar sunulan metodolojilerin görselleştirme ihtiyaçları için uygulanabilir olduğunu ve işlem yükünün önemsiz olduğunu ispatlamaktadır.

Anahtar Kelimeler: veri provenansı, veri görselleştirme, provenans verisi görselleştirme, insan bilgisayar etkileşimi

YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

1
INTRODUCTION

The term "provenance," which has French roots, describes the order in which a historical artifact has been in possession, custody, or location. The phrase is currently used in a variety of domains, including science, and computing, archaeology, paleontology, archives, manuscripts, printed books, and the circular economy, in ways that are comparable to how it was originally largely employed in regard to works of art [1].

By establishing, to the best of one's ability, an object or entity's later history. Particularly the sequences of its formal ownership, custody, and places of storage. The primary goal of tracing an object or entity's provenance is typically to provide contextual and circumstantial evidence for its original production or discovery. The method is particularly useful for assisting with object authentication. To achieve these goals, comparative methods, expert testimony, and the findings of scientific experiments may also be employed. But establishing provenance is mostly a question of documentation. In English, the phrase dates back to the 1780s. Conceptually, provenance is similar to the legal concept of chain of custody [1].

By preserving the data lineage in science and computing, provenance data records a data item's entire life cycle, from its point of origin through its final state. Systems for collecting data with provenance automatically record provenance. However, it's possible that their collection processes are flawed and have lost provenance notifications. As a result, provenance records may be incomplete, divided, or just wrong [2]. The analysis of provenance datasets is complicated by the incompleteness and inconsistent nature of provenance records if any exist. In many fields, provenance is required to guarantee reusability, repeatability, process optimization, fault tolerance, and performance estimation, including e-Science [3]. Investigating how scientists arrive at their conclusions also helps to validate and establish the reliability of evidence.

Recently, it has become crucial to analyze and explain massive amounts of data using

the five V factors: variety, volume, velocity, variability, and value. Particularly, the amount of data is growing significantly in the fields of e-Science and social media. In the big data era, organizations all over the world are producing more data on a regular basis. The information on the origin and life cycle of the data also grows quickly as the amount of data does. As a result, big provenance data is a term used by numerous scholars to describe data provenance [4]. Big provenance data analysis and visualization are needed for many applications [5].

Additionally, the amount of the actual data size may be exceeded by the metadata volume gathered by provenance monitoring. In order to conduct historical, current, and future research, it is crucial to collect, record, and mine data provenance. Data visualization is a significant topic that offers a variety of methods for creating an intuitive interpretation of data, frequently through the use of effective visualization tools [6]. Many professionals view data visualization as visual communication. By using diagrammatic displays that make use of data attributes, variables, and information units, several visual communication strategies develop processes that enhance visual data representation. By making difficult data more accessible, understandable, and useable, effective visualization can aid users in the analysis and evaluation of data [7]. To obtain a complete perspective of the data and discover the data's information, data visualization must be properly incorporated into the application. Numerous strategies for creating an ingrained data interpretation are proposed by data visualization [8]. A key step toward increasing the availability, transparency, and usability of complex data for effective perception maximization is efficient data visualization [9] [10].

Significant studies have identified a number of visualization strategies to fulfill provenance visualization needs and shed light on provenance visualization studies, particularly in the e-Science domain for the interpretation of workflow data. These researches served as the foundation for this study's analysis of provenance visualization requirements and realization of visualization methodologies.

The origin of the study data has a significant impact on the validity and dependability of scientific conclusions. As shown in the literature review section, much progress has already been achieved in obtaining tools and standards, but more collaborations and implementations are required for spreading frameworks and methods as much as possible while facilitating insights in a user-friendly way. This study focuses on the problem of enabling the required visualization processes utilizing a powerful visualization tool and removing knowledge graphs representing provenance data from scientific workflows.

1.1 Literature Review

The majority of research on provenance visualization focuses on a graphical display, graphic comparison methods, and small-size graphics. In order to evaluate provenance traces in a general fashion, PROV-O-Viz presents a web-based visualization tool for PROV-based provenance data supplied by various sources. A study[11] suggests a visualization framework to aid in the utilization of provenance data for scientific studies using ontologies and techniques for networks in the E-SECO scientific ecosystem[12].

In an interactive visualization project, the comic strips were produced from recorded provenance graphs based on quantified self-workflows[13]. Scientists from five different fields can find and explain map flaws with the use of the visualization tool Probe-It[14]. Due to its emphasis on data visualization techniques like summarizing and contrasting provenance data graphs, this study differs from earlier studies in this field. Additionally, the suggested visualization techniques have no direct connection to any one workflow system.

With the help of Kunde's important study[15], the requirements for the method, involvement, time frame, outcomes, relationship, interpretation, and comparison of provenance visualization are presented. Chen satisfies Kunde's list of requirements with customized layout designs and a number of visualization techniques that make it easier to visualize network provenance in the OPM model[16]. By offering OPM visualization for network data analysis, Chen collected e-Science provenance data and put various visualization techniques into practice. Chen also offers a visual comparison using an advanced direct classification algorithm (DCA) that compares two OPM provenance nodes to see how similar they are for the OPM model. By suggesting PROV-O model visualization strategies including summarizing and contrasting charts, this study varies from Chen's work. The DCA algorithm is expanded for comparing two separate provenance graphics, similar to Chen's work. We also employ the model specification expression capabilities of PROV-O. For the sensitive contents and constraints, an abstraction of provenance data is suggested as a user-friendly tool in a study[17]. We provide a center-oriented clustering algorithm for PROV-O graphs in order to summarize graphs. Additionally, the Cytoscape visualization framework includes the PROV-O provenance graphic visualization plug-in.

Significant research has been done on managing metadata for e-Science workflow[18] [19] [20] and information service management. Additionally, a number of studies about provenance usage in a number of domains, including social media and e-Science, have been noted [21] [22]. Numerous works on modeling metadata in

various fields are also available [22]. The proposed study differs from others in that it places a strong emphasis on provenance metadata visualization. A study[23] focuses on streaming applications of diverse active users and establishing temporary data streams in IoT applications on the reusability of common data streams in the smart city cyber infrastructure rather than reusing workflows and outputs after execution. In our work, the PROV Ontology data model is implemented using a similar DAG model (PROV-O). In order to comprehend the data, a study[24] proposes geospatial queries on data gathering for the web-based spatial metadata. We introduce backward and forward provenance querying to identify the dependencies in the workflow with the aid of visualization features. Backward querying on PROV-O metadata is simplified to obtain the source and results of the incursion.

Provenance mining uses a graphic database and processing system prototype with capabilities including streaming provenance data import, and current/past status queries on provenance data using Quine tool in a study[25]. Even though the Scala language has a query language, our system can query using higher-level, approachable HMI query features. User-inspired filters are presented in a study[21] to help cope with enormous volumes of provenance traces by reducing the provenance amount. Processes for the backward provenance segment comprise agents for producing output data using OPM documents as well as input and intermediate data. In our research, we enhance Chen's backward dependency technique (BDM-Backward Dependency Matrix) to boost the speed of backward dependency queries for PROV-O provenance data.

In order to evaluate the usefulness of our PROV-O visualization approaches, we gave users access to our satisfaction questionnaires, asking them to rate the system as a whole on a Likert scale for each topic. The 10 GB provenance database produced in a study[26] includes real-world workflows using Indiana University's WORKEM emulator, as well as recognized flawed patterns. The OPM to PROV-O data generation process is one of the main procedures in this research because the W3C PROV-O provenance model is the language of our provenance research environment.Provenance of the NASA National Snow and Ice Data Center Distributed Active Archive Center archive is utilized in a study to provide provenance data in an OPM ontology model. These archives are distributed on a daily, weekly, and monthly basis and are acquired by NASA's Aqua satellite.

One of the datasets used in this investigation is AMSR-Eprovenance data that Indiana University instrumented. Our research creates an AMSR-E dataset that is W3C PROV-O compliant rather than using the OPM data model. In order to collaboratively create novel provenance analysis-related methodologies, the current study on provenance

data analysis and user interaction is mentioned[27]. Our preliminary findings were published in a conference paper[2] in 2018 and our findings regarding the visualization of streaming provenance were published in 2020[28]. Numerous research on massive metadata processing on distributed data storage platforms have been conducted in a variety of fields, including social media[29] [30], weather forecasting[31], and information systems[32].

These studies primarily concentrate on developing scalable, high-performance systems with minimal processing overheads. For the goal of effective visualization, we are interested in studying a lot of provenance data in our study. In contrast to earlier studies, we concentrate on visualization strategies that would improve scientists' comprehension of huge provenance data. Studies on handling provenance data for many domains, such as social computing[33], mobile computing[34], geographic information systems[35], and earthquake prediction[36], have been conducted. We concentrate on provenance produced in e-science workflows for the goal of visualization, which is different from prior research.

The design and construction of web service-based visualization services for geographic information systems is the topic of some studies on visualization services in the literature[37]. In contrast to preceding research, this study focuses on data provenance visualization. Large data sets are difficult to manage and analyze. By utilizing bitmap indexing and parallelization, Yildiz et al.[38] investigated a unique type of search capability called membership query to address the issues. As we concentrate on visualization skills in this study, we leave the difficulties in indexing and querying the large body of provenance data for future research.

Here we presented brief background information on the literature but a comprehensive literature review is given in this thesis study on data provenance visualization in Section 3. In addition Section 4 includes a literature review on the data visualization methodologies and Section 5 has a literature review on the usability study as well.

1.2 Objective Of The Thesis

The main goal of this thesis study is to provide novel visualization support for data provenance. We identify the following motivations and limitations in the existing approaches in data provenance visualization studies.

The need for systems that enable end users to evaluate the data life cycle and take appropriate action has grown as a result of data provenance. Global data volume is

growing daily, especially in the Big Data era. As data volumes grow, data lifespan and origin metadata also outlive themselves. Users may become overloaded and have trouble understanding and uncovering the meaning of the data due to the data's vast volume and complexity [39]. Accordingly, there is an emerging need to extract meaningful abstractions from big amount of data. Many trends and patterns of the raw data would be easily accessible through the accurately produced brief statements of the main points to ease analysis and understanding [40].

Provenance is a type of metadata that shows information about the ownership, authority, and history of data from its origin to its final state. When it comes to establishing trust in the data and its validity, lineage comparison of two sets of data with similar processing is regarded as a crucial function. Comparative analysis is important for troubleshooting errors or unexpected situations as well [39].

Provenance metadata includes precedence relations because it is lineage data that is positioned on the time axis. By examining the root causes and intermediary artifacts, geospatial queries can be used to understand an event's underlying causes. Another necessity for the metadata analysis is the query in the forward direction in order to assess the potential impacts of an unexpected occurrence in the future. To check the sources and results of the incursion and identify the dependencies in a flow of metadata, querying in both directions is required [40].

The conventional visualization approaches involve offline processing and storage of provenance data, which consumes a lot of resources that are not essential. With today's changing needs, the enormous volume of data produced over time can also produce tremendous amounts of provenance metadata. This size may be too large to store and analyze. It is possible to ease the analysis of data arriving in time by making efficient use of resources by streaming visualization of the provenance entities within a predetermined time period. Another prerequisite is distributed display of streaming data, particularly for metadata analysis in IoT systems [2].

The literature does not provide a thorough enough literature review of methodology and approaches for data provenance visualization. Numerous studies addressed a literature survey, however, they were carried out as minor SLR studies with topic-specific research objectives in some of the major provenance visualization studies [41] [42]. One of the main topics of research in metadata analysis, provenance visualization still requires a thorough Systematic Literature Review (SLR) to provide guidance to the researchers for the study of data provenance visualization in the literature.

The effectiveness of provenance visualization techniques that enable metadata

analysis by end users must be assessed. These approaches are created based on user requirements, and users should be the ones who assess the approaches' applicability. In order to evaluate the system by users using the tasks provided, to gauge the system's effectiveness, and to direct future studies, comprehensive usability studies on metadata visualization methodologies are crucial [43].

1.3 Hypothesis

With the light of the given objective and the motivations given in the Section 1.2, the hypothesis of this study is that it is possible to design and implement a set of user-friendly, interactive and interoperable provenance data visualization methodologies to enable end user better understanding of provenance data by summarization, similarity-differences analysis, visualization of provanance data stream and dependency querying in provenance data for the application use domains.

1.4 Statement of Research Problem

In the light of previously mentioned motivations, our research problem is, "What are the visualization methodologies that can enable end-users to better analyze and understand provenance data?". In order to address this research problem we identify the following research questions based on the motivations that we defined in the previous section:

- How can we design and implement a user-centric,interactive and interoperable provenance visualization tool including visualization approaches?
- How can we design and implement a visualization methodology that can group together data provenance to help end-users better understand the data?
- How can we design and implement a visualization methodology that can help end-users identify the differences and similarities between the two provenance data?
- How can we design and implement a visualization methodology that can help end-user to understand the dependencies in provenance data?
- How can we design and implement a visualization methodology that can help end-user to analyze streaming provenance data?

- How can we conduct a comprehensive literature review investigating the significant studies on Data Provenance Visualization in literature with a rigorous process?
- How can we conduct a comprehensive usability study research on provenance visualization tools to assess usability from the end user's point of view?

1.5 Thesis Scope

In order to define the scope of this thesis we outline the requirements of the desired visualization methodologies and also outline the application use cases.

1.5.1 Provenance Visualization Requirements

Data on provenance can be exceedingly big and dense. By alone, lineage records or provenance annotations offer a chance to record further details about the execution or production of data. A system should be able to provide incremental loading for reading provenance graphs with and without annotations in XML format in order to provide visualization over huge graphs.

A layout is an important part of data provenance visualization since it aids in user comprehension.User navigation models, HCI principles, and the nature of provenance data is to be taken into consideration in a visualization research. This may rely on the enhancement of current layouts and the development of new layouts.The visualization approach should generate customized layouts that can enable user to display provenance data in layers. Example layouts to help user understanding the relations in time shown in Figure 1.1, Figure 1.2, and Figure 1.3.

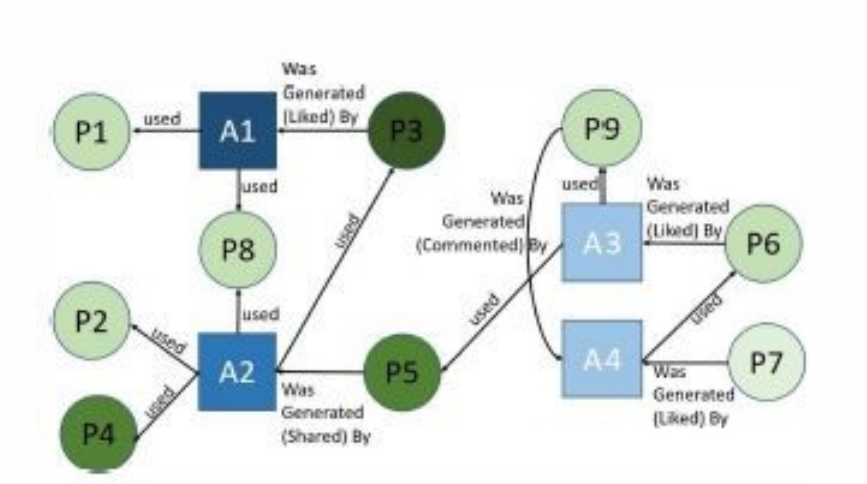


Figure 1.1 Sample Process-Artifact Relations in time

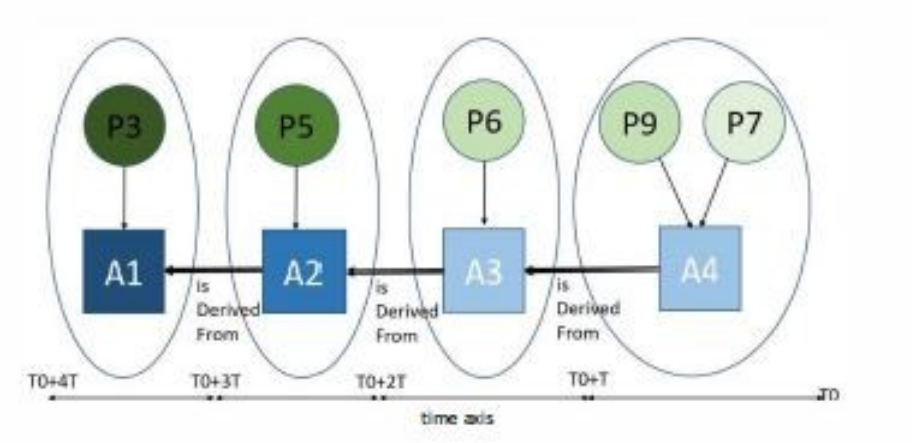


Figure 1.2 Sample Individual Process-Artifact Relations in time

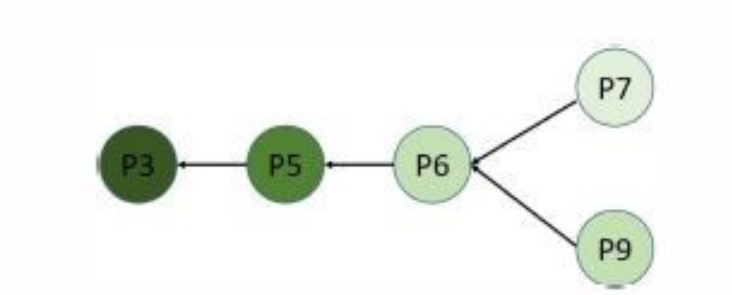


Figure 1.3 Sample Lineage Process order

The user is the main driver of the visualization systems, and the user can boost usability and efficiency by taking advantage of the proposed methodologies. Existing visualization styles are issues to investigate in order to create an efficient visualization. Data provenance visualization systems should provide user-centric visualization by following Human-Computer Interaction (HCI) principles. They capture users' attention and attraction and increase usability.

Users become overloaded and have difficulties understanding and uncovering the meaning of the data due to the metadata's vast volume and complexity. This needs brief explanations of main points. Meaningful abstractions should be provided by summarizing provenance graphs in order to generate brief statements. And also eliminating redundant items increase understanding and analysis capabilities.

Finding similarities and differences between same workflow's two e-science graphs in different time is important in visual mining for spotting anomalies. The visualization approach should be able to succeed in comparing provenance graphs to determine differences and similarities.

Many internet services and decision-making applications use patterns gleaned from

massive amounts of data more and more in the data-centric era. To satisfy these objectives, huge scale stream processing systems in particular analyze flowing data as infinite streams. In provenance visualization systems, it should be able to provide stream visualization. Identifying the input data leads the development of any output data item created in time accordingly.

When a data sequence is executed, metadata can have some unexpected or problematic results which are displayed as output. The user would like to investigate the input or intermediate artifacts that caused that condition. Similarly, examining the potential nodes that could be impacted by the subsequent unexpected or undesirable occurrences is crucial for the provenance analysis. Backward and forward dependency querying should be provided by the data provenance visualization systems to interpret dependencies.

1.5.2 Application Use Domains

Concepts for provenance visualization are regarded as challenging to create. It is important to take care of keeping the link between the particular requirements of an application domain and the interpretation of data provenance when developing a general visualization technique. The main goal of provenance visualization should be to create several visualization concepts and assess them in light of specific needs. The provenance community supports the advancement of various visualization methods and assesses them for potential application fields. We outline potential application domains and their specific requirements for data provenance visualization in the following.

1.5.2.1 E-Science Domain

Description: E-Science is computational, science-intensive data found in highly distributed network environments. The capture, modeling, and preservation of provenance data are the main topics of research into e-Science provenance. Because their systems are continuously functioning for lengthy periods of time to facilitate scientific experimentation, e-Science has access to very huge volumes of data with increasing complexity [16].

E-Science can involve a variety of fields, including biology, physics, and chemistry. Any experiment that is conducted in a digital environment must follow a set of stages before a final result, such as a digital output, a photograph, or a text file, may be produced. E-Science workflows are typically lengthy, multi-step processes that take a long time to complete.

A workflow, which is used in many e-Science disciplines, describes all of the processing phases in an analysis. Process management makes the connections between different workflow steps and their computing needs plain so that the whole workflow can be run independently. For workflow execution, workflow management systems like Kepler are facilitated in a computational environment and metadata is generated as output.

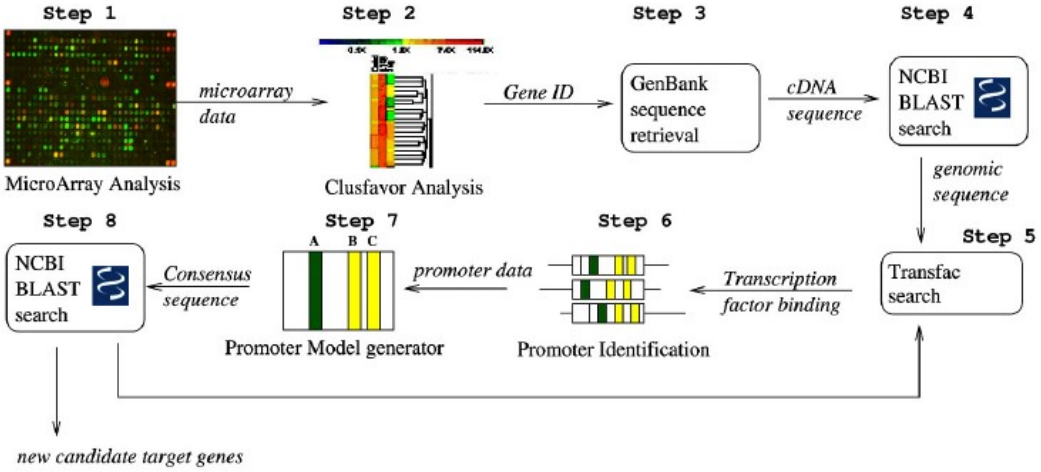


Figure 1.4 Conceptual ("napkin drawing") view of the Promoter Identification Workflow (PIW) [44]

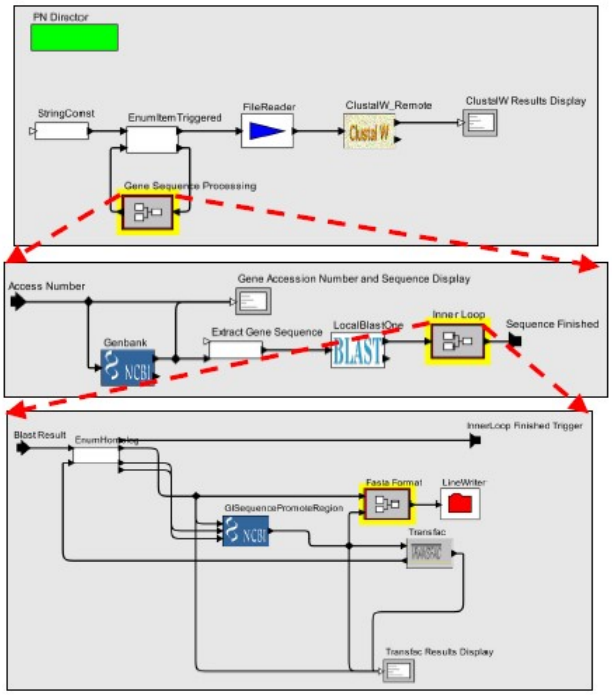


Figure 1.5 PIW implemented in Kepler [44]

Figure 1.4 shows an illustration of a Kepler workflow for DNA sequence processing with 8 steps inside. Kepler includes user interfaces to execute workflow steps for the

scientists. Sometimes a workflow consists of sub-workflows that operate concurrently as shown in Figure 1.5. A workflow is executed inside the Kepler system and generates provenance data as metadata regarding its execution. For monitoring and analyzing the effectiveness of this workflow, a retrospective analysis of this study is crucial. Workflow management systems like Kepler is a workspace management systems that supports workflow executions but usually don't provide visualization support for the analysis of this metadata.

Requirements: Unexpected or incorrect results are the main issue while implementing these workflow processes. The scientist should be able to examine the process and go step by step to determine which processes are responsible for those undesirable or unexpected results. Therefore, methodologies are needed to analyze the log files created during these processes and to visualize them accordingly.

Scientific workflows including sub-workflows such as Figure 1.4 and Figure 1.5 are being run over the course of hours or days in a distributed manner, producing vast quantities of provenance data that are overwhelming for the scientist to process. Since provenance graphs are huge and complex, it is frequently necessary to produce summaries from complex graphs in order to facilitate analysis and comprehension on brief statements.

The scientists run the workflows several times and generate provenance data accordingly for each run. In order to analyze incorrect or unexpected results, two provenance data generated by the same workflow at different times should be compared. By checking differences and similarities in each step one can identify the problematic nodes of workflows in each step. In addition, the scientists should be able to affect the comparison by adjusting properties' weight and threshold values to customize comparisons.

In order to analyze the root causes of unexpected or problematic results, the provenance output should be analyzed to discover input and intermediate nodes causing the defect in a workflow. Similarly, workflow nodes affected by problematic nodes should be investigated by examining the selected node's intermediate and output nodes. Backward and forward dependency functionalities are needed for querying and debugging in a scientific workflow.

Scientific investigations produce enormous provenance data sizes that are difficult to store, analyze, and visualize over the course of lengthy periods of time and on a broad scale. Online visualization of streams in a defined time frame is a developing need to analyze scientific artifacts generated in workflows. By providing replay functionalities, scientific experiments can be reproduced in backward and forward directions.

1.5.2.2 Social Network Domain

Description: Provenance data associated with a social media statement can help dispel rumors, clarify opinions, and confirm facts. Taking advantage of social media features, research in this nascent field spearheads the search for a way to provide provenance data to social media users, thus leveraging social media itself by mining it for the provenance data.

Social workflows represent an abstract view of the various social patterns observed on Twitter. It can be understood, visualized, and represented in different formats and, thus, analysis can also be conducted on it.

Users of Twitter tend to use a set of undeclared rules when tweeting, replying, or re-tweeting in order to control their social engagement as shown in a sample workflow in Figure 1.16. For example, a tweet that starts with a mention is considered part of a conversation with low visibility that other users will most likely not see, or using a hashtag as a keyword engages in the general social context [29].

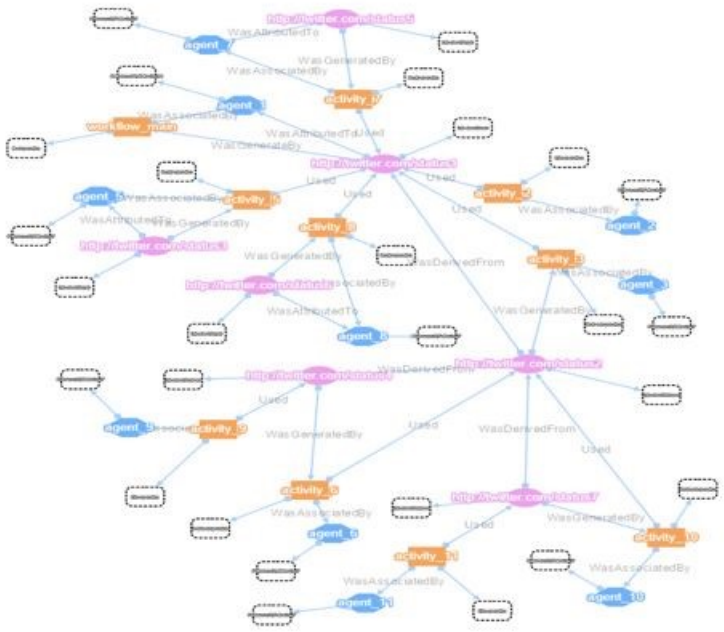


Figure 1.6 An example social media provenance data[29]

In social media, actions are taken on a social media artifact such as a tweet such as posting, retweeting, reposting, and mentioning. The retrospective analysis on the actions is a requirement to discover insights and to realize misinformation. A data provenance visualization approach is a need to conduct basic and advanced analysis of social media data.

Requirements: Social networks have transformed marketing and, as this post shows, their popularity is still growing in our latest global social media statistics research summary for 2022. Networks vary in popularity with different demographics and they're still evolving. Research by Global WebIndex that we reference in this article shows that globally more than half of the world, 4.70 billion people around the world now use social media and 227 million new users have come online within the last 12 months.

It is important that the huge amount of social media data be grouped in such a way that meaningful summaries can be drawn and abstract views are presented to the user. In a social workflow, a past story of social media elements such as a tweet is an issue of analysis to discover its history on it such as posts, retweets, and like actions in the backward direction. In the same way, future actions of a tweet data at a time is an issue of metadata analysis, for example, a trend analysis of a top tweet is another need for action analysis. Hence backward and forward dependency analysis is to be supported with a visualization approach.

Streaming visualization of social media is another need to monitor social media artifacts and the users' actions on them in the time axis. For example, actions such as reply, retweet, and mention analysis in the arrival order can be observed online by checking the properties using the memory resources effectively. And with the replay option, stored provenance data can be reproduced in time using time labels accordingly.

1.5.3 Thesis Contributions

The creation of provenance visualization is challenging due to the need for provenance data interpretation. Maintaining the connection between the data and its application is one of the most persistent issues. In this thesis study, we design and implement a number of different visualization methodologies to address the previously defined requirements. We outline the intended contributions of this study as in the following.

- Creating an effective and user-centric provenance visualization methodology for visualizing provenance data [39] [45].
- Designing and creating interoperable data provenance visualization methods are achieved either by enhancing the ones that already exist or by introducing new methods [2] [2].
- Creating a visualization prototype that is built based on the requirements that we outlined in Section 1.3.1 [2].

- Designing and implementing a data comparison visualization methodology to find the differences and similarities of given graphs [2] [39].
- Designing and implementing a scalable summarization methodology that can enable users to better understand the data provenance [2] [45].
- Designing and implementing a visualization methodology that can visualize streaming provenance data [45].
- Designing and implementing a visualization methodology for exploring dependencies in provenance data [45].
- Conducting an SLR study on existing data provenance visualization studies [46].
- Conducting a comprehensive usability study on a given provenance visualization approach [43].

1.6 Organization of Thesis

The organization of this thesis study as in following. Chapter 2 introduces the fundamental concepts including data provenance definitions such as ontology model details and tool definitions as Chapter 3 provides detailed background information on provenance visualization by referencing previously conducted SLR study including experimental studies and contributions. Chapter 4 focuses on the proposed data provenance approaches' details by giving use cases, design, implementation, and experimental results for each. While chapter 5 explains comprehensive usability study details that take place to assess the proposed visualization approaches' success by defining the usability study process and experimental results. Chapter 6 provides the conclusions. The evaluations and discussions for the particular contributions are provided in the corresponding chapters.

2
FUNDAMENTAL CONCEPTS

2.1 Data Provenance

The term "data provenance" refers to the life cycle of the data, which includes its generation, origin, and distribution. This process is viewed as the data object's recorded past. How an object was made, changed, spread, and brought to its present state. A data object's credibility can be deduced by looking at the data provenance. Data provenance has previously been investigated to verify physical artworks, literary works, and manuscripts. Today, data provenance is receiving a lot of attention in the digital data world, particularly in the e-science domain, where it is essential for debugging, auditing, and assessing the accuracy and dependability of data. Complex ecosystems in the online world of today increase the pervasiveness of data. For academic groups in databases, workflows, and distributed systems, handling provenance data is becoming more important.

A data record with provenance information identifies the people, organizations, and processes that contributed to the creation of the data or the release of a portion of it [47]. To evaluate data quality and trust, improve reproducibility, and to effectively debug complex data systems, understanding the provenance and lineage of the data is a crucial task. For past, present, and future inquiries, data provenance creating, recording, and analyzing activities are crucial steps [15]. Data provenance and metadata are frequently linked to improving understanding of the data and processing of specifics. Provenance monitoring is highly regarded as an important function for visualization systems because it allows scientists to view data in a derived view and to make observations about reliability and data quality, both of which are crucial for reproducibility and trust in scientific outputs [19].

To have a complete data picture, to uncover insights, and to make difficult data more open, accessible, and available, data visualization provides a variety of ways and processes that assist data preparation naturally. It should be flawlessly integrated into the analysis tools. Particularly interactive data visualization facilitates expert

and non-expert users in a substantial way, bringing data exploration and analytical outcomes. Through explanation and exploration visualization techniques that let users engage with the data, visualization aids in the provenance discovery process. To facilitate data mining and improve consumers' view of big data, visualization designers suggest a wide range of functions for provenance metadata, including grouping, repeatability, and collaboration.

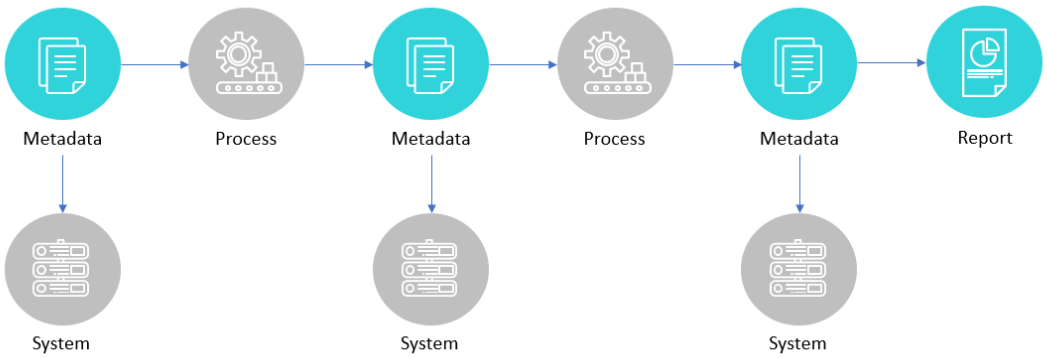


Figure 2.1 Data lineage illustration

Data provenance and computational reproducibility are crucial to achieving improvements in research quality. Research from the past demonstrates how challenging it was to accurately duplicate computational experiments. Furthermore, the Large Data phenomenon makes this process more difficult than it was in the past, and recent research has focused on the benefits and challenges of provenance, particularly in the context of big data [48].

The W3C defines provenance as data on the organizations, processes, and individuals involved in creating a given piece of data or object. This data can be used to make judgments about the object's caliber, dependability, or credibility. The PROV Family of Documents offers a model, related serializations, and other ancillary definitions to allow the interoperable exchange of provenance data in heterogeneous settings like the Web. An overview of this document family is provided in this document [47].

For provenance architectures, the concept of provenance is essential. Other terminologies used in earlier studies to describe this concept are below:

Data Quality: Lineage information based on the source data and the transformations can be utilized for data quality estimate and data reliability. Lineage also offers evidence for data elaboration.
Audit Trail: Data generation can be improved through resource consumption analysis and error detection using provenance data, which supports traceability on audit trails.

Replication: Provenance can be used to ensure maintenance and replication when repeating data derivations.
Informational: Generally, the lineage data is utilized for query-based data discovery on metadata and data interpretation.

The amount of fine-grained provenance data can be overwhelming for end-user and data analysts. Along with supportive queries over provenance, this problem can be solved by generating proper, interactive visualizations to empower the end-user to explore and analyze provenance information [49].

2.1.1 Provenance Taxonomy

To support the requirements for data provenance in each domain, many methods have been used. In a survey study on provenance [50], a conceptual taxonomy is provided along with brief analyses of their benefits and drawbacks. In Figure 1, the taxonomy is summarized. The five major sections are covered one at a time. Each section is defined in the following briefly.

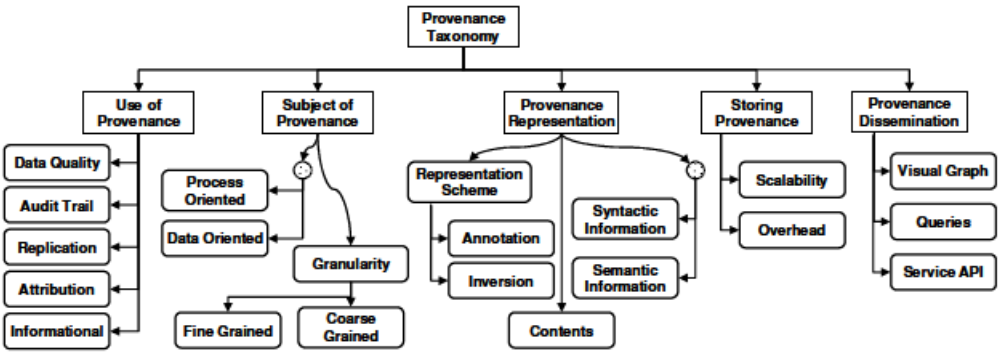


Figure 2.2 Provenance Taxonomy

Use of the provenance section in Figure is identified by referencing Goble's study [47] as mentioned in the previous section such as Data Quality and Replication. In the data processing system, provenance data can be gathered about various resources at various levels of detail. The lineage of the provenance data may be immediately available or inferred indirectly. A data-oriented model, which we refer to as an explicit model, collects lineage metadata, particularly for the data product. In contrast, in a process-oriented, or indirect, paradigm, the deriving processes serve as the main objects for which provenance is gathered, and the input and output data products of these processes are examined to ascertain the data provenance. The level of detail at which provenance is gathered determines how relevant information is in a particular field [50].

Provenance data can be represented using a variety of methods, some of which are

dependent on the underlying data processing technology. The expense of recording provenance and the depth of its application depends on how it is represented. Annotations or inversion are used in the two main methods for conveying provenance information. Since each profession has different needs, it is difficult for a suitable metadata standard to develop because there isn't one already in place for lineage representation across disciplines. Present-day provenance systems that make use of annotations frequently employ XML to express the lineage data. Some use domain ontologies in languages like RDF and OWL to capture semantic data inside provenance. The concepts and connections used in the provenance data are stated in ontologies in a precise manner [50].

If the data is fine-grained and rich in provenance information, the provenance information itself may eventually grow to be larger than the data it describes. Therefore, the provenance metadata's scalability depends on how it is stored. It is possible for provenance to be closely related to the data it describes, to be housed in the same data storage system, or even to be included directly in the data file itself, as is the case with the NASA Flexible Image Transport System headers. Such methods can make it simpler to preserve provenance integrity, but also make it more difficult to publish and search solely the provenance. Provenance may be kept separately or in conjunction with other metadata. Fees associated with provenance management include storage and collecting costs. To save on storage costs, provenance data that isn't utilized as often can be archived, or provenance data for data that is used frequently can be retained using a demand-supply model based on usefulness [50].

A system must provide extensive and varied access options in order for provenance to be used. A derivation graph that users can browse and inspect is a typical method of distributing provenance data. Users can also look for datasets using provenance metadata, for example, to find all datasets produced by a specific procedure. These query results can automatically feed input datasets for a workflow at runtime if semantic provenance information is present. Datasets' derivation histories can be used to replicate data at different locations or to update data that has become outdated as a result of changes made to its forebears. Additionally, provenance retrieval APIs can let users develop their own usage patterns [50].

2.2 Provenance Ontology Model

It is not now possible to expect all systems to adopt a single method of describing and gathering provenance. Instead, it is essential to take into account a fundamental provenance data model to enable the translation of domain- and application-specific

representations into a similar data and exchange model for interchangeability. Because of this, applications that need to understand provenance in systems can import, process, and reason about it. As a result, heterogeneous systems can export their provenance into such a core data model.

Any semantic technology is built on ontologies. A set of specifications linked to a concept is the official definition of an ontology. Ontologies are tools that many researchers use to share and reuse data. Relationships between identifiers can be simply expressed using ontologies. They are similar to knowledge representation systems in many ways.

2.2.1 The World Wide Web Consortium (W3C) Working Group

W3C [47] is a community that develops Web standards and guidelines to help it carry out its goal. The purpose of the Provenance Working Group is to facilitate the widespread distribution and utilization of provenance data for Web documents, data, and resources within the W3C community. A language for transferring provenance data among applications will be defined in the Working Group's W3C Recommendations.

The PROV Ontology and accompanying Notes were successfully published, and on June 19th, 2013, this organization was shut down. Consult the PROV page of the Semantic Web Wiki, which is updated by the community, for more details, relevant tools, etc. The W3C defines provenance as a record that lists the individuals, groups, organizations, and processes involved in creating, shaping, or distributing a piece of information or a physical object. A piece of data or object's provenance includes information about the entities, actions, and individuals involved in creating it, and it can be used to make judgments about its caliber, dependability, or trustworthiness. To enable the interoperable exchange of provenance information in heterogeneous settings like the Web, the PROV Family of Documents offers a model, appropriate serializations, and other supporting definitions. This document gives a summary of this document's family.

The W3C Provenance Working Group sought to express unique meta data on the data by providing a view of the full revision chain of the data, asking who played what role in its creation (author, reviewer, etc.), asking which parts of integrated data come from which original data and under what process, and asking which vocabularies/ontologies/rules were used to generate some portions of the data [47] ?

2.2.2 W3C PROV Family

This is a member of the PROV family of documents, a collection of documents outlining various requirements for realizing the goal of interoperable provenance information exchange in heterogeneous settings like the Web. Figure 2.3 provides examples of these documents, which are also given below.

- PROV-OVERVIEW (Note), an overview of the PROV family of documents
- PROV-PRIMER (Note), a primer for the PROV data model [51];
- PROV-O (Recommendation), the PROV ontology, an OWL2 ontology allowing the mapping of the PROV data model to RDF [52];
- PROV-DM (Recommendation), the PROV data model for provenance [53];
- PROV-N (Recommendation), a notation for provenance aimed at human consumption [54];
- PROV-CONSTRAINTS (Recommendation), a set of constraints applying to the PROV data model [55];
- PROV-XML (Note), an XML schema for the PROV data model [56];
- PROV-AQ (Note), mechanisms for accessing and querying provenance [57];
- PROV-DICTIONARY (Note) introduces a specific type of collection, consisting of key-entity pairs [58];
- PROV-DC (Note) provides a mapping between PROV-O and Dublin Core Terms [59];
- PROV-SEM (Note), a declarative specification in terms of first-order logic of the PROV data model [60];
- PROV-LINKS (Note) introduces a mechanism to link across bundles [61].

2.2.3 W3C PROV-O Ontology Model

The PROV Ontology (PROV-O), which represents the PROV Data Model (PROV-DM) using the OWL2 Web Ontology Language (OWL2) [OWL2-OVERVIEW], is chosen as the data provenance standard for this study. It offers a collection of classes, limitations, and attributes that can be used to express and exchange provenance data produced by various systems and in various settings. In order to model provenance information for

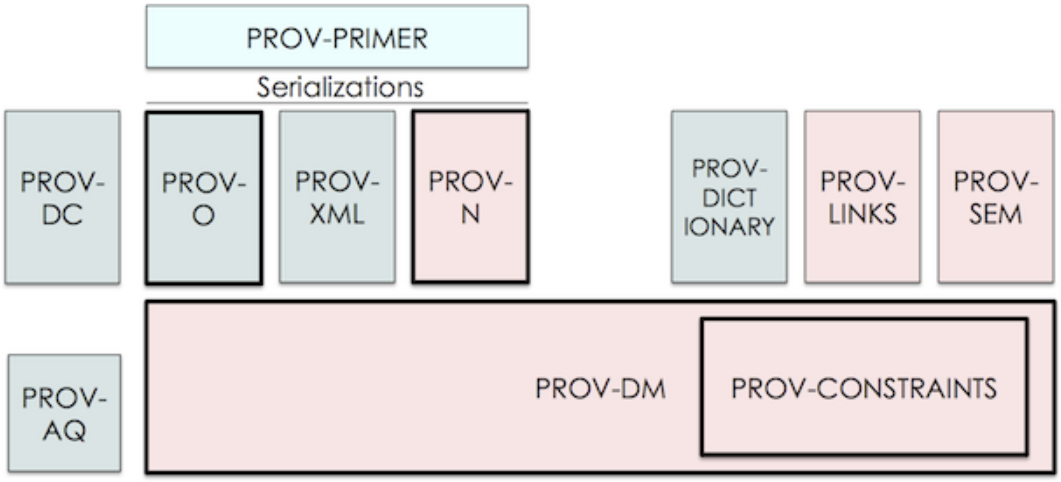


Figure 2.3 W3C PROV Family

various applications and domains, it can also be adapted to produce additional classes and characteristics. The PROV Document Overview, which should be read before other PROV documents, summarizes the overall situation of PROV [47]. Simple, introductory provenance descriptions can be made using the few classes and properties in the Starting Point category. The rest of PROV-O is built upon these three classes:

- An **prov:Entity** is a thing that can be physical, digital, intellectual, or of another kind with certain established characteristics; entities can be real or hypothetical.
- An **prov:Activity** is an action that takes place over time and is performed on or on behalf of entities; it can involve consuming, processing, transforming, altering, moving, using, or creating entities.
- An **prov:Agent** is something that is somehow accountable for an action occurring, for an entity existing, or for the actions of another agent. The attributes depicted in the accompanying image are used by the three basic classes to relate to one another and to themselves.

Activities can use and produce a variety of Entities over their lifecycle and at specific times in time (specified using characteristics prov:startedAtTime and prov:endedAtTime, respectively) (described with prov:used and prov:wasGeneratedBy, respectively). For instance, a blog writing task might take a certain dataset and produce a bar chart. One can build provenance chains that include Activities and Entities by stating usage and generation.

Additionally, without explicitly stating the start and finish times of the activities, we can indicate that an Activity prov:wasInformedBy another Activity to convey

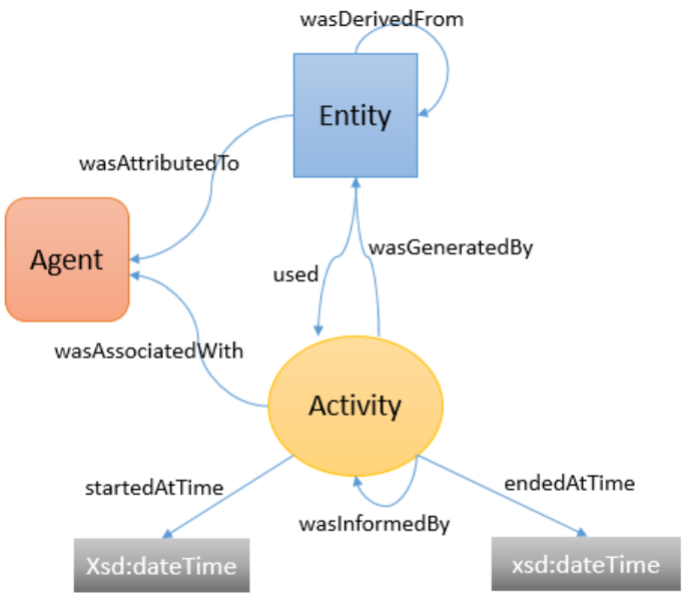


Figure 2.4 W3C PROV-O components

some dependence information. A prov:wasInformedBy relationship between Activities demonstrates that the informed Activity utilized an Entity that was created by the informing Activity, even while the Entity itself is unknowable or of minor significance. Since only Activities are allowed in provenance chains, the prov:wasInformedBy property enables this.

The prov:wasDerivedFrom property can be used to create provenance chains that are made up only of Entities. An entity is transformed into another through a derivation. For instance, we can state that the bar chart prov:wasDerivedFrom the dataset if the Activity that generated it is unknown or not of interest. The fixed characteristics of an Entity that are useful for a specific application can be described using arbitrary RDF properties (for example, the file size and format of the dataset, or the aspect ratio of the bar chart).

While provenance chains between Activities and Entities can be created using the properties prov:used, prov:wasGeneratedBy, prov:wasInformedBy, and prov:wasDerivedFrom, Agents may also be given responsibility for any Activity or Entity inside a provenance chain. The values prov:wasAssociatedWith and prov:wasAttributedTo, respectively, reflect the extent to which an Agent is accountable for an Activity or an Entity. Agents may also be held accountable for the deeds of other Agents. In this instance of delegation, the influencing Agent provided that they were acting on behalf of another Agent who is likewise accountable for the influenced Activity or Entity. Prov:type and prov:label are expressed, respectively, by

the properties rdf:type and rdfs:label.

2.3 Data Provenance Visualization

A literature survey is conducted to realize the current state of the visualization methodologies and tools to assist end-user towards reproducible research. Kunde, whose work is regarded as novel study in research on provenance visualization, describes the user needs for the technology from a higher perspective in the innovative study [15].

Abstract user needs for provenance visualization are derived by Kunde et al. include: 1) process: the inspection is focused on the order in which the process steps are performed; 2) outcomes: the focal point of the user's view is the interaction's final or intermediate outcomes; 3) relationship: it's crucial to consider how actors or interactions are related; 4) timetable: it's necessary to be aware of the time; 5) participation: it's crucial that participants are accurate; 6) compare: contrasting two subjects reveals their differences; 7) interpretation: a unique visualization display based on the end-particular user's query.

2.4 Systematic Literature Review (SLR)

As a rigorous methodology, SLR is a crucial tool for synthesizing references in a certain field of study. Investigating all pertinent data related to predetermined study questions is the goal. Kitchenham proposed a guideline to execute SLR for Software Engineering by reference from medical research guidelines after the first SLR for the medical domain was presented in [62]. In this important study, SLR is viewed as a crucial activity for locating, evaluating, and analyzing potential material in relation to a certain set of questions. While the individual studies that are being published for this review are referred to as main studies, the SLR study is referred to as secondary research.

SLR is more scientific than a basic literature review because of its meticulous approach. In literature, SLR studies are carried out for a variety of reasons, including summarizing the available evidence regarding a technology, identifying gaps in the current research area to be able to suggest additional investigations, and providing a common background to support new research activities properly.

2.5 HCI Design

User interface and experience are the main focuses of HCI. It's crucial to create interactive computer systems that are effective, efficient, simple, and pleasant. because it is advantageous to those who use the interactive system. There are several guidelines to abide by when creating interactive systems.

A number of evaluators are given an interface design and asked to evaluate a user interface application in order to get a sense of the positive and negative aspects of an HCI. This type of usability study is known as heuristic evaluation. Significant studies have been done in the literature to specify the rules for carrying out heuristic evaluations. In this study, we analyze our suggested visualization application using leading heuristics principles, which are dubbed "heuristics" introduced by Nielsen [63], and then build user interface accordingly. The following is a brief explanation of Nielsen's ten usability heuristics, which are more like guidelines than usability principles.

1- Visibility of system status (Feedback): Users are kept up to date on the state of the visualization process by the system, which provides pertinent feedback in a timely manner.

2- Metaphor Usage- matching between system and real world: Instead of using system-specific terms, the system makes use of the user's natural language, including phrases and concepts that are recognizable to the user.

3- User Control ve Freedom: Users may accidentally activate some interface features and search for an emergency exit to get out of the undesirable state without continuing the conversation.

4- Consistency and standards: Users may be left wondering if different actions and names can result in the same awful design. By taking this into account, the visualization user interface (UI) functionality was developed to have unique functional groups for a specific purpose, such as clustering and comparison, to prevent user confusion.

5- Error Prevention: First and foremost, the application's UI functionalities are cleverly created to guard against potential mistakes. However, problems with the visualization parameters, inconsistent data, and other issues could still occur.

6- Recognition rather than Recall: It's crucial to keep the user's memory burden as low as possible by making objects, actions, and options obvious and apparent. The user prefers not to retain knowledge from two conversations, instead choosing to follow

directions that are clear and appropriate for the action.

7- Flexibilty and Efficiency of use: There might be choices for rookie and experienced users, allowing users to customize regular actions, to speed up interactions for expert users and minimize processes.

8- Aesthetic and Minimum Design: To keep the design minimal, we avoid adding unnecessary dialog boxes and UI elements. Functionality that satisfies the basic requirements for visualization is included in UI design.

9- Help Users Recognize, Diagnose and Recover from Errors: User comments and error messages are written in straightforward language that defines the issue.

10- Help and Documentation: Although it is actually preferable if the application can be used without a user manual or documentation, having either can make it simple for users to be guided whenever they require.

2.6 Usability Testing

Usability occurs independently of computer systems, and its name has become more widely used as Human-Computer Interaction has developed into an interdisciplinary field. The usefulness of a developed product is a key measure of its quality. A product with a high level of usability enables a user to use it effectively, achieve their goals after using it, and feel fulfilled by doing so.

The process of evaluating a product or service using actual users is called usability testing. Typically, during a test, participants will try to complete common tasks as observers watch, listen, and take notes. This study aims to evaluate the usability of the product, collect qualitative and quantitative data, and gauge participant satisfaction. A well-planned test design, participant recruiting, analysis, and result reporting are necessary for an effective usability test.

Usability testing is advantageous because it enables the design and development teams to identify problems before they are programmed. The less expensive a problem is in terms of staff time and potential schedule disruption, the sooner it is identified and fixed.

The usability testing process generally entails the following steps: determining whether participants can complete the required tasks successfully, estimating the time required to complete the listed tasks, gauging participant satisfaction with the application or website, figuring out what changes are required to improve user

performance and happiness, and assessing the effectiveness to see if it meets the usability goals.

2.6.1 System Usability Scale

In this study, usability is evaluated using the system Usability Scale (SUS) method. SUS, developed by John Brooke in 1986, gives a "quick and dirty" trustworthy tool for evaluating usability [64].)It allows to the evaluation of a wide variety of products or applications.

It is possible to talk about certain major usability metric categories using SUS. The effectiveness of a system should be evaluated by users' capacity to complete tasks with it, efficiency should be evaluated by the amount of resources utilized to complete activities, and satisfaction should be evaluated by users' subjective reactions to using it, states ISO 9241-11.

For respondents to evaluate the system and gauge their satisfaction with it, Brooke established an SU scale in his innovative study. Prior to now, we presented our satisfaction survey's findings in two groups. In addition to these, SUS score survey was asked to the test participants to assess the overall system for general level satisfaction utilizing Likert Scale values. Likert Scale values are given in Table 2.1 for each of the survey questions illustrated in Figure 2.5.

Table 2.1 Likert Scale values

Value	Point
Strongly Disagree	**1** point
Disagree	**2** point
Neutral	**3** point
Agree	**4** point
Strongly Agree	**5** point

2.6.2 SU Scale Evaluation:

Each score is added to the SU total value to create the SU scale value by subtracting 1 from the score for odd-numbered questions and subtracting their value from 5 for even-numbered questions in the SUS metrics and the SUS result is assessed according to the SUS categorization [65] shown in Figure 2.6.

	The System Usability Scale Standard Version	Strongly Disagree					Strongly Agree
			1	2	3	4	5
1	I think that I would like to use this system frequently.		O	O	O	O	O
2	I found the system unnecessarily complex.		O	O	O	O	O
3	I thought the system was easy to use.		O	O	O	O	O
4	I think that I would need the support of a technical person to be able to use this system.		O	O	O	O	O
5	I found the various functions in this system were well integrated.		O	O	O	O	O
6	I thought there was too much inconsistency in this system.		O	O	O	O	O
7	I would imagine that most people would learn to use this system very quickly.		O	O	O	O	O
8	I found the system very awkward to use.		O	O	O	O	O
9	I felt very confident using the system.		O	O	O	O	O
10	I needed to learn a lot of things before I could get going with this system.		O	O	O	O	O

Figure 2.5 SUS Assessment Question Set

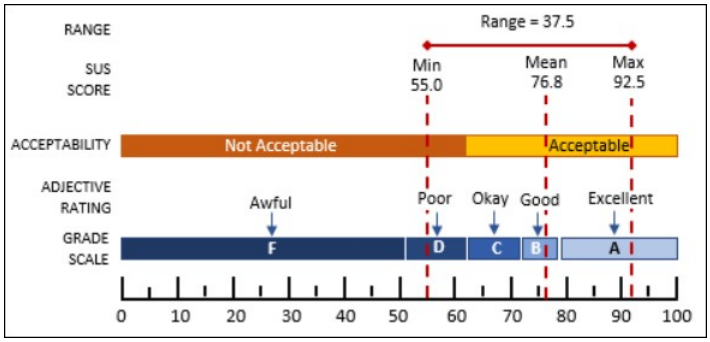


Figure 2.6 SUS Score Categorization

3
SYSTEMATIC LITERATURE REVIEW

This work is being regarded as a reference for examining specific research on data provenance visualization methodologies and technologies because it proposes a significant systematic review of the literature. Presenting a SLR to explain works that explicitly address data provenance visualization in general or in other areas, specifying visualization approaches and techniques, is the common objective of this research [46]. A thorough screening was enabled with primary studies in the literature by adhering to a strict procedure. In this study, we conduct an SLR study on data provenance visualization in order to fill the gap by meeting the comprehensive literature review need in this area.

3.1 Overall Approach

In this SLR study, the guideline steps below are derived from studies, it concludes with three main steps, Review Planning, Review Process, and Review Reporting [62]. Review Planning is the description of the requirement for a literature review. The review Process includes the step for identifying the review research, selecting reference studies, studying quality evaluation, data extraction, screening, and data interpretation while Review Reporting is reporting the results of the usability study by outlining the purpose behind the survey, the background itself clarifies the need for the assessment. Section 3.8 defines the Related Works section.

3.2 Systematic Literature Review on Data Provenance Visualization

In this section, we are going to go into the details of the SLR study. We will first start with the research questions followed by the detailed descriptions of the search string and data sources definition, study selection and data extraction, and study selection and data extraction results. Then the discussions for each research question will be

presented in detail with comprehensive tables. Threats to validity topics and related works and contributions will be given to conclude the section at the final step.

3.2.1 Research Questions

Which studies in the literature focus on data provenance visualization to suit the analysis needs in domains? is the primary question posed during the SLR investigation? A series of research questions are established to address this core subject, as shown in Table 3.1.

Table 3.1 Research Questions.

ID	Research Question	Main Focus
R01	How is the distribution of the publications in years?	Distribution
R02	What was the venue of publications?	Venue
R03	Which domains data provenance is proposed/used?	Domain
R04	What provenance ontology/standards are used in data visualization?	Interoperability
R05	Which visualization approaches are needed to meet visualization needs?	Purpose
R06	How is end-user validation handled in studies?	Usability

As seen in the third column, each defined question specifically targets a research topic set. The topic of this SLR study for knowledge synthesis is primary studies.

3.2.2 Search String and Data Sources

To create the standard search terms and choose the databases for the literature papers, a literature analysis was carried out. By conducting advanced specific searches on well-known digital libraries to collect the primary studies, we specifically focus on data provenance visualization.

In order to construct the search string, the most popular terms were first determined. The 'OR' operator for searching for synonyms of search terms and the 'AND' operator for grouping synonyms together are present in all of the published digital library sources, supporting searches on digital papers. As a result, the following search term was employed in this SLR study: "DATA PROVENANCE VISUALIZATION" OR "PROVENANCE VISUALIZATION"

Table 3.2 provides examples of the cited digital data sources. Since each data source has its own distinctive search patterns, the search sentences must be modified while retaining their essential components for each new search activity.

Table 3.2 Digital Data Sources

Source	URL
IEEE Explore	**h**ttps://ieeexplore.ieee.org/Xplore
ACM Digital	**h**ttps://dl.acm.org
Web Of Science	**h**ttps://www.isiknowledge.com/
Springer	**h**ttps://link.springer.com/
Science Direct	**h**ttps://www.sciencedirect.com/
Scopus	**h**ttps://www.scopus.com/

3.2.3 Study Selection and Data Extraction

The World Wide Web Consortium (W3C) announced the first official draft for the reference provenance data specification known as PROV Document Family [47] in December 2012, hence publications published before that time were disregarded. Additionally, some inclusion and exclusion criteria for choosing pertinent publications to be reviewed in this SLR study were introduced below.

Inclusion Criteria

IC1- Automatically identified research articles and papers focusing on data provenance visualization in electronic data sources.

Exclusion Criteria

ExC1- Duplication papers
ExC2- Secondary papers
ExC3- Papers are not written in English
ExC4- Review Articles,Surveys,Book Chapters, Tutorials,Reports

According to the type supported by the search engine, the research papers were exported from the electronic databases in BIBLEX, RIS, or CSV format, and prepared using free Jabref and Zotero tools. The SLR Process was properly supported by the results that were fed into the stART program. The Free Of Charge tool stART is designed to carry out SLR procedures in accordance with the methods outlined by Kitchenham [62].

The literature citations were arranged, and SLR processes including data collection and extraction were controlled. The SLR results were ultimately exported to a spreadsheet.

3.2.4 Data Extraction Results

Examining and reading all the chosen data to help the literature study is not a practical method. The following fields were used to categorize the initial data: title, writers, publication database, kind, and year of publication. Figure 3.1 provides a detailed explanation of our SLR process.

The literature search was done in July 2022, and it produced a total of 2923 papers. The papers were initially rejected based on the previously stated Exclusion Criteria. 305 manuscripts were chosen as candidates for the data extraction phase after each paper was individually examined in Phase 1 of the Figure by reading the reading tile, the keywords, and the abstract fields. Phase 2 involved a thorough study of each of the chosen papers, and 18 publications were ultimately extracted to address the SLR research questions listed in Table 3.1 in this phase.

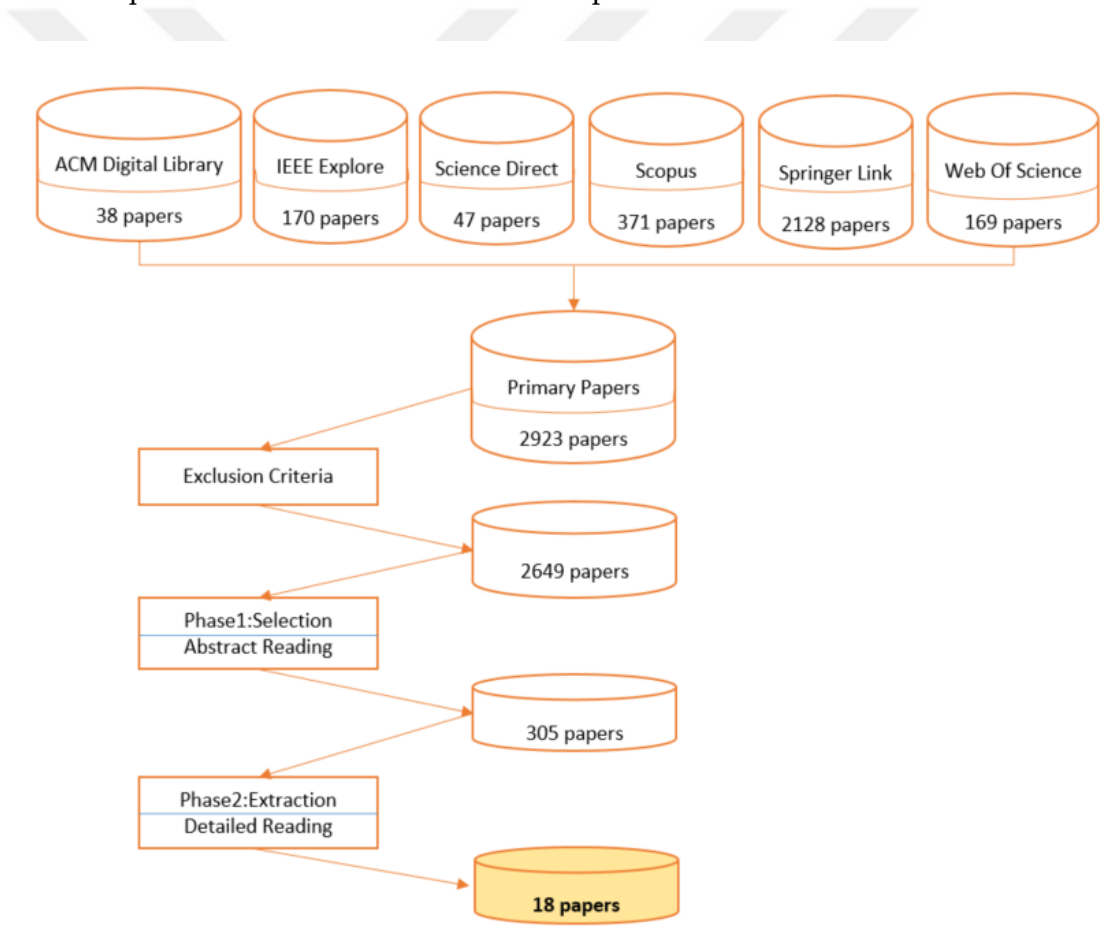


Figure 3.1 SLR process steps

The remaining 18 studies are the focus of the current review of the literature, and they are listed in Table 3 along with their kind (CN: conference, JR: journal) and citation data. The following headings are used to summarize the papers in accordance with how well they respond to the research questions that drove this SLR study.

Table 3.3 Categorical Data and Reviewed Studies Reference mapping

Study	Reference	Type	Year
SP-01	[66]	CN	2016
SP-02	[45]	JR	2021
SP-03	[67]	CN	2020
SP-04	[68]	JR	2013
SP-05	[69]	CN	2017
SP-06	[70]	JR	2016
SP-07	[71]	JR	2019
SP-08	[40]	CN	2012
SP-09	[42]	CN	2021
SP-10	[72]	JR	2013
SP-11	[73]	CN	2016
SP-12	[74]	CN	2021
SP-13	[75]	JR	2019
SP-14	[13]	CN	2017
SP-15	[11]	CN	2015
SP-16	[76]	CN	2021
SP-17	[77]	JR	2022
SP-18	[6]	JR	2018

3.2.5 Discussion of Research Questions

This section provides a thorough assessment of the literature while addressing the particular research questions we listed in Table 3.1. The distribution of publications throughout the year will be provided in the first place, and the publication venue will be covered in the second. Following that, the study domains and ontology models will be given. Finally, proposed visualization strategies and study usability validation will be discussed.

3.2.5.1 RQ1:How is the distribution of the publications by year?

The studies are spread out between the years of 2012 and 2022, and the distribution appears to be balanced. As shown in Figure 3.2, we discovered that the years 2016 and 2021 had the most papers, three and four, respectively. There are fewer papers from the years 2012, 2015, 2017, 2020, and 2021, which is one. On the other side, no papers related to our search topic were published in 2014.

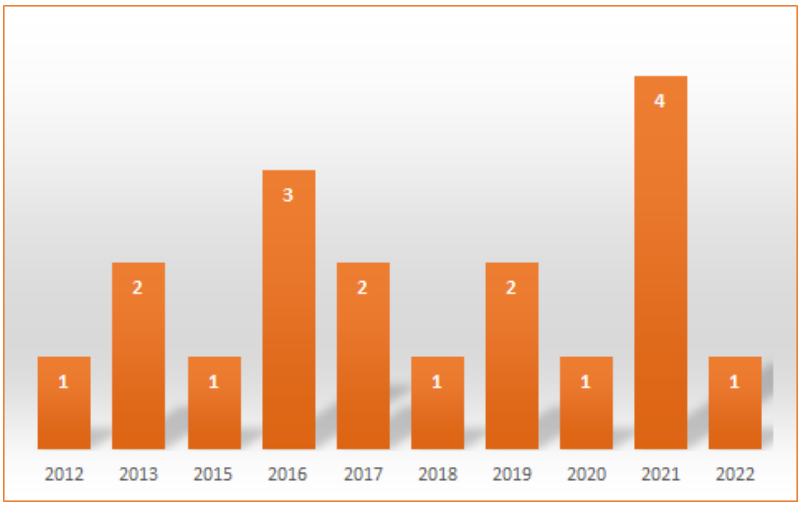


Figure 3.2 Publications in years

3.2.5.2 RQ2:What was the venue of publications?

Clarifying the data sources for the published studies is the aim of the research topic. Ten papers were published in conferences, while eight papers were published in journals, for a close to the equal number of publications. These conference papers show that they are not yet developed and contain sufficient information for journal publication. Table 3.4 below shows the list of journals and conferences where the SLR articles were published.

Table 3.4 Journals and Conferences That Reviewed Papers Published Under

Journal/Conference Name	Digital Database
IEEE Int. Conference on Trust, Security and Privacy in Computing	IEEE Explore
Concurrency and Computation: Practice and Experience	Wiley Online
IEEE Visualization Conference (VIS)	IEEE Explore
IEEE Transactions on Geoscience and Remote Sensing	IEEE Explore
Biennial Conference on Innovative Data Systems Research	Conference Resource
Computer Graphics Forum	Wiley Online
Workshop on Massive Data Analytics on Scalable Sys. (DataMASS)	IEEE Explore
International Conference Information Visualisation (IV))	IEEE Explore
IEEE Transactions on Visualization and Computer Graphics	IEEE Explore
International Provenance Annotation Workshop	Springer Link
Working Conference on Software Visualization (VISSOFT)	IEEE Explore
IEEE Computer Graphics and Applications	IEEE Explore
USENIX Workshop on the Theory and Practice of Provenance	ACM Digital
International Provenance and Annotation Workshop	Springer Link
ACM IKDD CODS and COMAD	ACM Digital
Knowledge and Information Systems	Springer Link

The majority of the three papers published in IEEE Transactions on Visualization and Computer Graphics journals are related to IEEE journals, as can be seen in the table. This is an important conclusion for our study since it will point researchers to the venues where articles concerning data provenance visualization are most frequently published. The other publishing platforms are Wiley Online Library, ACM Digital, and Springer Link.

3.2.5.3 RQ3:Which domains data provenance is proposed/used?

In order to assess the domains and use cases where provenance visualization needs are addressed inside the studies, this research question is posed. The majority of studies claim that while specific use-cases are illustrated to draw attention to suggested visualization aspects, provenance visualization methodologies are examined for generic provenance visualization reasons. Table 3.5 lists the selected studies' studied topics and use-case regions. As shown in the table, network management and E-Science domains are the ones in which data visualization is studied as an emerging need.

Table 3.5 Study and Domain/Use Case Relation

Study ID	**Domain or Use case Area**
SP-01	Cloud Computing, Security
SP-02	E-Science
SP-03	Web application
SP-04	GeoScience, E-Science, Cyber Security
SP-05, SP-13	Data management, data wrangling
SP-06	Biomedical, E-Science
SP-07	data retrieval, biomedical
SP-08, SP-09, SP-10, SP-15, SP-17	Network, File Systems
SP-11	Georeferencing
SP-12	Software Development
SP-14	Health Data
SP-16	Social Data
SP-18	Time Series

3.2.5.4 RQ4:What provenance ontologies/standards are used in data visualization?

This research question seeks to evaluate the languages or data model specifications used in the studies to express provenance data. For the purpose of interoperability, the studies were analyzed. It has been noted that most studies concentrate on W3C provenance models, whereas other studies specify bespoke provenance languages or employ JSON as a data format. Table 3.6 below has the specifics.

Table 3.6 Provenance data model specification proposed in the studies

Study ID	Provenance Data Model Specification
SP-02, SP-11, SP-12, SP-14, SP-15, SP-17	W3C PROV-O Family
SP-04	RDF, OWL
SP-03, SP-05, SP-06, SP-13, SP-16, SP-18	Customized data model
SP-01, SP-07	JSON,CSV
SP-08	W3C OPM
SP-09	RDF
SP-10	Provenance-Aware Storage System, PASS

The W3C provenance ontology models were the first ones selected by the investigations, as shown in Table 3.6, since the standardization satisfies the provenance visualization standards established by Kunde et al. [15]. An acceptable provenance representation model satisfies the first three requirements, which are concerned with the process, results, and relationships.

3.2.5.5 RQ5:Which visualization approaches are studied to meet visualization needs?

The research question is asked to investigate visualization attributes and studied visualization approaches in the reviewed papers. Visualization features are compared in logical groups in order to support intelligibility, so-called visualization features, and interactive visualization attributes. Each is handled in a subsection with a detailed SLR table. In each table, the studies that have X-mark in respective fields are the studies that indicate this feature is proposed in these studies, blank fields are either not proposed or not specified.

Visualization Features: The visual aspects of provenance visualization are evolving from traditional procedures to new technologies and approaches. This is still an unexplored subject that researchers are currently studying. The table given in Table 3.7 examines and presents in-depth the visualization aspects identified as functional demands in the publications under examination.

HCI principles are increasingly crucial in these investigations since data visualization is directly tied to human perception. For instance, by introducing Gestalt principles into the system at the system design stage, Garae et al. [66] attempted to boost usability efficiency.

Interactive Visualization Components: This section examines interactive visualization techniques and the information about them that is offered in the reviewed publications. The reviewed studies all provide simple navigation for

Table 3.7 Visualization feature comparison

	Basic Features					Operational	Quality)			Implementation	
ID	Name	Visual* Representation.	Temporal	Workflow Support	Abstract (AV) Multiple (MV) Views	Collaboration (CO) Web Based (WB) Configurable Parameters (CF) Annotation (AN) Action Recovery(RC)	Gather Insight	Reproduction	User Centric Design	Open Src	Tools GUI Libs
SP1	UVisP	TM,HM,BD,FD	X		MV				X		D3
SP2	N/A	GV	X	X	AV	AN		X	X	X	Cytoscape
SP3	Trrack	GV			AV	CO,WB,CF,AN,RC	X	X	X	X	Javascript
SP4	GePWProv	WM		X	AV,MV	WB		X			GeoPWT, GeoPOM
SP5	DVMS	PD,BC	X	X	AV,MV	CF,AN,RC	X				
SP6	Avocado	GV		X	AV,MV	CO		X		X	Dagre,D3
SP7	KnowledgePearls	GV,PD,BC	X		AV,MV	CO,AN	X			X	Gapminder
SP8	N/A	GV,FD		X	AV,MV	AN			X		Cytoscape
SP9	MGExplorer	GV,MT,CV,BC	X		AV,MV	CO,AN,RC					
SP10	InProv	RTL	X		VM				X		Processing
SP11	ProvViewer	GV	X	X	AV,MV	CF,RC				X	Jung
SP12	N/A	GV,BC,PD,CM			AV,MV	WB	X			X	GIT-LAB2PROV
SP13	DQProv Explorer	GV,BC	X	X	AV,BC	CO,AN,RC	X				OpenRefine
SP14	ProvComics	GV	X	X	AV	WB,AN			X	X	ProvStore JQuery
SP15	Prov-O-Viz	GV	X		AV	WB					
SP16	N/A	GV			AV					X	
SP17	Visionary	GV	X	X	AV,MV	WB	X			X	
SP18	SomFlow	GR,CM,DM FD,MT,HM.	X	X	AV,MV	CF,AN	X		X		Java 2D

* GV: Graph View,CV: Circular View, FD: Force Direct graph.
RTL: Radial based Tree Layout
BD: Bubble Diagram, PD: Plot Diagram
BC: Bar Chart, GR: Grid, MT: Matrix
WM: Web Map, TM:Tree Map, HM:Heat Map, CM: Color Map

activities and entity tracking using metadata. The interaction elements and features that were made available to the end user in the trials were thoroughly reviewed and detailed in Table 3.8.

Table 3.8 Interactive visualization Components comparison

ID	Clustering/ Grouping	Comparison	Graph Merge	Details on Demand	Stream Visualization	Filter (F) Highlight (H) Sort (SO) Search (SE) Zoom (Z)	Query by Definition	Query by Example	History Retrieving
SP1				X				X	X
SP2	X	X		X	X	S,F,SE,Z		X	X
SP3	X							X	X
SP4		X		X		SE,F,H			X
SP5	X				X	F,H	X		
SP6	X	X				F,H,Z,SE			X
SP7	X	X		X		SE,F,H		X	X
SP8	X	X				Z			X
SP9	X	X		X		F	X		X
SP10	X			X		F,SE,Z			X
SP11	X		X	X		F,Z,H			
SP12						F			X
SP13		X		X		F,H			X
SP14	X			X					
SP15						F,H			
SP16	X								X
SP17		X				F,SE,H			
SP18	X	X		X		F,Z,H			

3.2.5.6 RQ6: How is end-user validation handled in studies?

The goal of to ask this question is to investigate the user validation methodology of the studies. Since the end-user is at the heart of such systems, it is important that the system is evaluated by the user / user group. Some of the studies conducted a usability study for the validation and the rest planned to do a user evaluation process in the future.

In the following studies [13, 43, 45, 66, 72, 75] planned usability studies carried out by user groups and the evaluations were presented accordingly. Especially in the studies [45] addresses a SUS usability metric applied for e-Science domain using e-Science documents [43] and in [43, 45, 72], a comprehensive usability assessment steps were defined ans users were asked to answer several questions related to systems' usability, using Likert Scale defined in the studies. According to the responses, the systems' validation were assessed using average and mean values.

3.2.6 Threats to Validity and Limitations

The following categories of factors could have an impact on the research's findings. Some papers might have been wrongly included or excluded from the SLR study.To prevent this, the rigorous research technique was reviewed among the researchers to ensure a shared understanding with the aim of eliminating this threat, and inclusion/exclusion criteria were applied for study selection and data extraction. In addition a specialized SLR tool, StArt was facilitated to execute SLR process by integrating all bibliographic records into one location in StArt tool to evaluate primary studies gathered from all data sources.

In order to avoid missing any studies that might be relevant to this SLR, we conducted searches in all significant academic data sources. We preferred to query all fields (title, keywords, abstract, ...etc.) and full text if the search engine enabled it when conducting these searches. Additionally, when supported, we used wildcards to include all conceivable keyword variations; otherwise, we only included the majority of them in query strings.

Since domain-specific factors are mostly incorporated in the visualization process, not every component of data provenance visualization could be covered in detail by the study questions. In order to prevent this, we made an attempt to develop a broad range of research questions and sought to concentrate on factors that are domain-independent.

3.3 Related Work and Our Contributions

With a subset of issues about visualization features and methodologies, various research that addresses literature reviews or SLR studies on data provenance visualization has been done.

A survey on provenance data analysis strategies, including user-centric approaches and system-centric approaches such as adaptive systems, was presented by Xu et al. in their paper [27]. A three-stage process was established to conduct literature reviews by addressing papers published between 2009 and 2019 and posing why, what, how, and where queries to the chosen studies. The primary objective was to enhance provenance analysis advances through the multidisciplinary study of user interactions in the domains of artificial intelligence, machine learning, and HCI. **In this study we differ from this study by having more comprehensive digital data libraries and addressing a wider scope by asking research questions accordingly while they ask Why, What and How questions only. In addition, we conduct a well-formed SLR directly on the data provenance visualization systems by issuing more recent**

papers, between 2012-2022.

Suh et al. survey's study [41] looked at the visualization capabilities of simulation provenance systems in addition to the modeling, capturing, and querying elements for scientific simulations and workflows by categorizing systems into several relevant areas. **However our study addresses all relevant studies on data provenance visualization from all computational domains not restricted to scientific workflows. Furthermore, we investigated more visualization capabilities in groups while they look for just a few such as querying and reusability. And again we address papers between 2012-2022 while they cover till 2018.**

In the important study, [78] Oliveira et al. carried out a partial SLR investigation by including 10 chosen studies looking into several fundamental features, such as filtering, viewing, and searching, as well as the study itself to provide context and present the study's goal. Our study differs from this study by conducting a comprehensive SLR study while [78] addresses a minor SLR study as a literature review before proposing their data provenance visualization framework, Visionary. **Only functional type requirements are investigated while we address six different research question sets such as domain, distribution, and usability besides functional requirements in our comprehensive study.** Menin et al. compared data visualization technologies in [42] by examining characteristics like temporal, distributed, and relational properties of five chosen visualization tools. **However, they focused on only a few studies summarizing features according to the functions of their proposed tool, our in-depth survey includes a comparative analysis by evaluating all the papers between 2012-2022. Moreover, our study differs from this study by following a rigorous SLR process instead of a conventional literature survey by asking comprehensive research questions.**

Contributions: In our, SLR study, each research topic offers data provenance visualization options with a variety of general or niche visualization techniques. The corresponding SLR paper was published as a conference paper [46] and it investigates 18 relevant papers between 2012 and 2022.

This SLR can be regarded as the first thorough SLR study in this field, despite some literature reviews being undertaken as minor SLR studies in some of the major provenance visualization research as mentioned above. The research has demonstrated that provenance visualization is one of the most important areas of study to address the needs of metadata analysis, particularly in the e-Science domain. Since every study suggests future plans highlighting the various unanswered questions that still need research in this field, it appears that the area is still being investigated.

4
DATA PROVENANCE VISUALIZATION METHODOLOGIES

In this chapter, we discuss data provenance visualization methodologies that we propose in this thesis scope. We discuss the data visualization methodologies to address the motivations that we identified in Section 1.1. We also discuss the visualization methodologies in order to meet the requirements we outlined in Section 1.3.1. The methodologies that we discuss in this chapter provide answers to the research questions that were outlined in Section 1.2.

4.1 Overall Approach

In this section first we discuss detailed research questions that we want to answer related to our proposed visualization methodologies. Then we discuss the requirements and use cases that we outline in order to in designing the proposed visualization methodologies and in answering the research questions in Section 4.2.1.

4.1.1 Research Questions

We aim to address the following research questions either by designing and implementing the ways during the thesis study in order to meet provenance visualization requirements by the proposed visualization approaches.

Table 4.1 Research Questions

ID	Visualization Requirement
RQ1	How to design and implement an interoperable provenance visualization?
RQ2	How to design and implement user-centric visualization framework?
RQ3	How to provide detailed viewing of node properties?
RQ4	How to design and implement provenance graph summarization?
RQ5	How to examine scalability in summarization visual approach?
RQ6	How to design and implement comparison of the experiments/workflows?
RQ7	How to design and implement backward, forward dependency querying?
RQ8	How to provide streaming provenance data visualization?

4.1.2 Requirements and Use Cases

Within the thesis study, we investigated the user interface requirements of end-user to perform visualization methodology to meet the provenance visualization needs. In this section, we reiterate requirements that we previously outlined in Chapter 1 under Section 1.3.1 and we also describe the use cases that correspond to these requirements. In each use case diagram, we simply specify the expected behavior of the studied visualization methodology as use cases specifying inner use cases and the relations between them. Nielsen's heuristics mentioned in his novel study [63] are taken into account while designing the use case diagrams.

Interoperable Provenance Visualization: Our objective is to take advantage of well-defined web semantics, including data models and definitions, in order to facilitate the interoperability for provenance information interchange in visualization. A complete provenance data specification requires three fundamental ontology item types. Entity items include things, whether they are physical, digital, or other types of things; activity items include actions that use existing entities or create new ones; and agent items include things that participate in activities and bear responsibility.

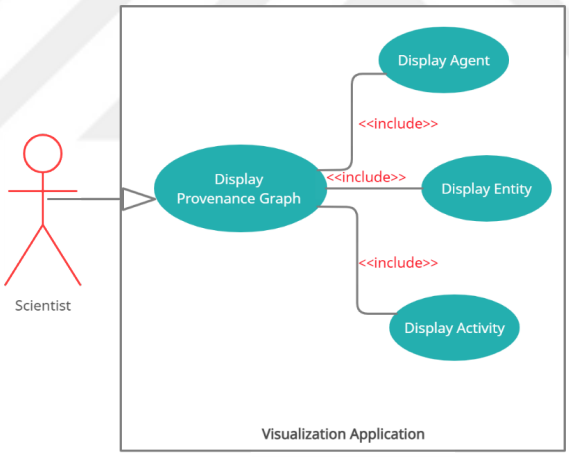


Figure 4.1 Use cases for interoperable ontology-based provenance visualization

The use case diagram in Figure 4.1 depicts the ontology components displayed from the user's perspective. As seen in the inner use cases Display Agent, Display Entity, and Display Activity, when a user attempts to display provenance data with Display Provenance Graph, the ontology components of the data, activity, entity, and agent items should be displayed seamlessly without requiring any additional user action.

User-Centric Provenance Visualization: To have a user-centric set of visualization approaches, we aim to provide the basic visualization features to the end-user to support user navigation properly. A basic view option allows scientists to view, highlight or focus on a node by changing the display. So the scientist can identify

most influential nodes in the graph network.

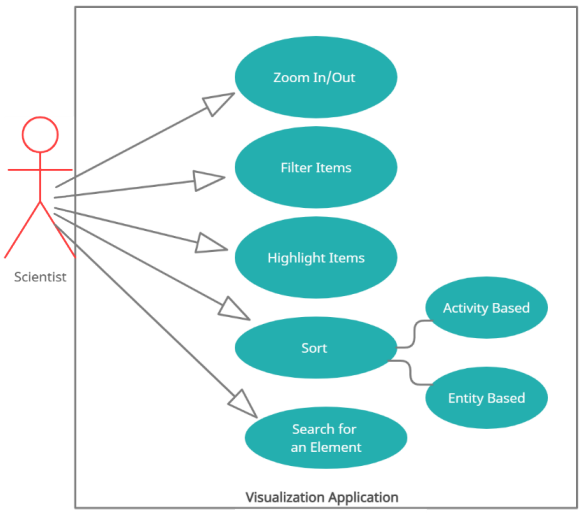


Figure 4.2 Use cases for user-centric provenance visualization

The detailed use case diagram is shown in Figure 4.2. The basic visualization features, zooming, filtering and highlighting an element, searching for an element should be supported by the visualization tool. In addition, for example e-Science data provenance is proposed in the scientific activities to mandate descriptive metadata from a rich set of data. Reusability and tracking of steps from the data derivation to the end values has a significant impact in such data intensive e-Science projects. Sorting of activities and entities in time and a detailed node inspections are another requirement will be addressed as a visualization feature of proposed framework.

Provenance Data Summarization: Considering with large-scale provenance data graphs it becomes crucial producing brief summaries to help better understand on data properties. Grouping of the large scale data into the particular set of items on their characteristics helping to create brief summaries for the user, so that the user can focus on the particular groups directly. Clustering is the task of grouping different data object is classified as similar objects. One groups means a cluster of data.Data sets are divided into different groups in the cluster analysis which is based on similarity. In this study, a property-based clustering is utilized to generate brief representation on node property similarities. Regardless of the clustering technique, user should be able to define the cluster count to dispatch the provenance data items to the corresponding clusters properly.

Making concise summaries to aid in a better understanding of data attributes is vital when dealing with large-scale provenance data graphs. In order for the user to immediately focus on the specific groupings, it is helpful to organize the enormous amounts of data into a specific set of objects based on their characteristics. The task

of grouping various data objects into categories of comparable objects is known as clustering in data mining. Based on similarity, data sets are split into various categories in the cluster analysis and one groups refers to a collection of similar data. In this study, a property-based clustering is issued to produce a concise representation of the similarities between node properties. No matter the clustering method, the user must be allowed to specify the cluster count in order to properly assign the provenance data items to the appropriate clusters.

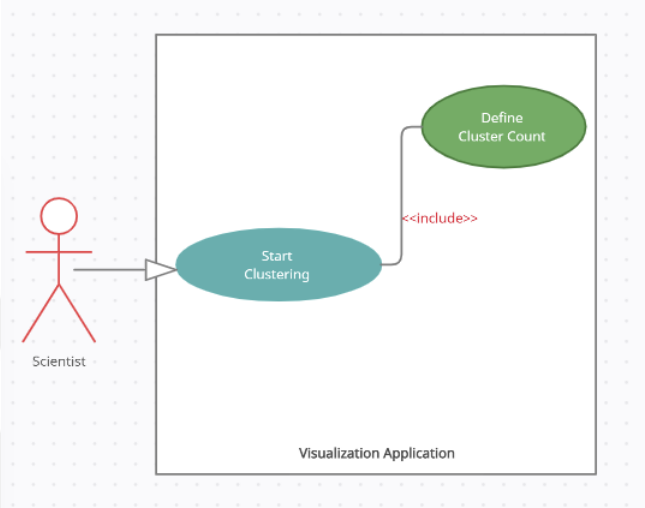


Figure 4.3 Use cases for provenance data summarization

Figure 4.3 shows the visualization system's inner use case for defining cluster count and its main use case for initiating clustering using the suggested user interface. The user can immediately evaluate the outcomes of clustering for various cluster count selections by having a cluster count configuration.

Comparison of the two provenance data: Identifying differences and similarities between two data graphs for the purpose of anomaly identification or debugging is one of the key responsibilities of data provenance analysis via data provenance visualization.

To meet the needs of users for comparison visualization, we present two separate comparison modes in this study: a basic comparison and an advanced comparison. Basic comparison only takes into account the label values of the compared nodes, but advanced comparison in advanced mode also takes into account node and edge properties. The use cases for the system that the user interface design will provide are shown in Figure 4.4. The user first imports two graphs to be compared to the visualization system and then mode selection is employed via the user interface. The specifications of the node and edge property weights and the threshold values that will

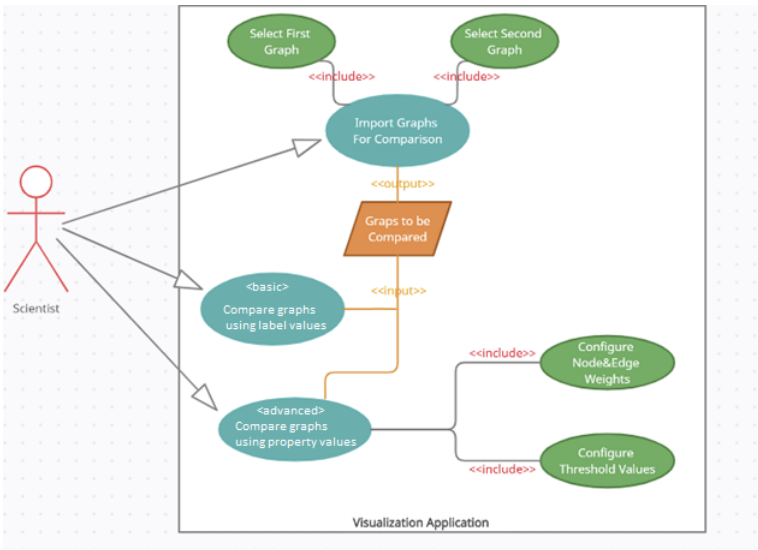


Figure 4.4 Use cases for provenance data comparison

be applied during the comparison process are included in two more inner use cases that are part of the advanced mode selection.

Provenance data dependency querying: Provenance data analysis and inquiry must, as specified in the requirements section of Section 1.3.1, connect PROV-O nodes and their connections in a cause-and-effect manner. To evaluate the dependencies for agent, entity, and activity ontology items, we provide two dependency querying solution in both backward and forward directions in this visualization approach.

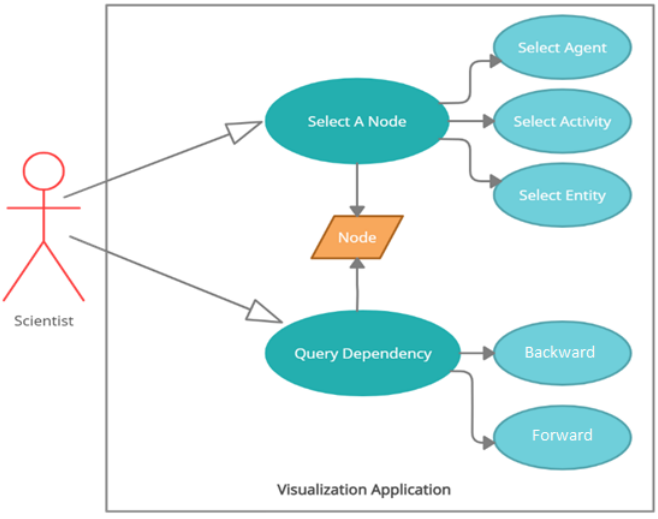


Figure 4.5 Use cases for data provenance dependency querying

The Figure 4.5 illustrates the use cases in the system for applying the dependency querying methodologies in both directions. First the node selection is done from one of the agent, artifact, entity nodes and then the forward or backward dependency

direction is chosen to execute dependency querying.

Provenance data streaming visualization: The conventional provenance data visualization approaches involve offline processing and storage of provenance data, which consumes a lot of resources that are not essential. Large-scale, long-running scientific investigations can produce a significant amount of provenance data that is difficult to retain. We discuss a streaming visualization where entity, agent and activity items of provenance data are represented with a finite amount of memory within visualization frame and in the sequence in which they come.

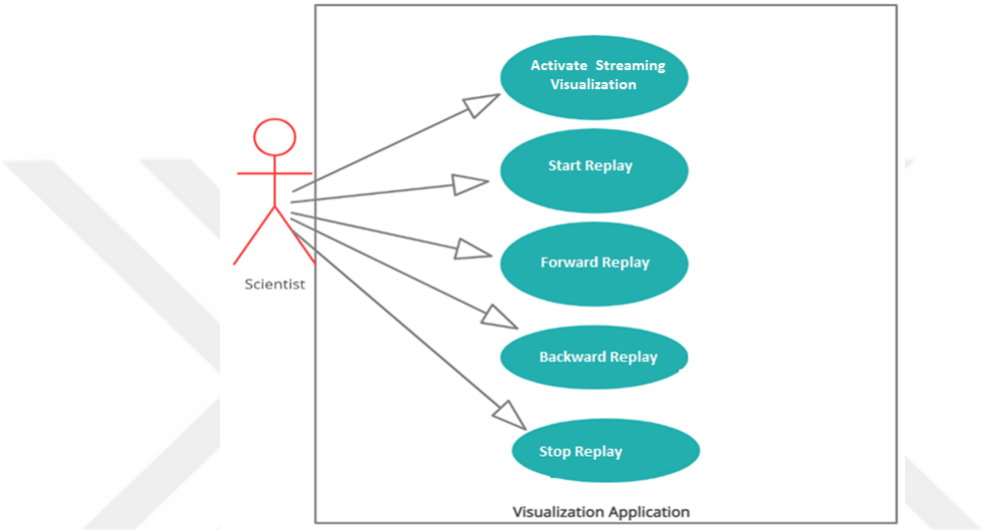


Figure 4.6 Use cases for streaming provenance data visualization

Figure 4.6 displays the use cases for streaming visualizations after importing the provenance data into the visualization system. The agent, entity, and activity components are visualized at the sequence of their arrival inside a time period with Start Replay after the streaming visualization is enabled using Activate Streaming Visualization. Following that, forward and backward replay in time can be seen for further analysis and replaying can be turned off with the Stop Replay use case.

4.2 Data Provenance Visualization Methodologies

In this section, we explain the proposed visualization approaches including design and implementation details. First we will talk about proposed visualization methodologies, second we will talk about the datasets that we used in order to prove that the proposed methodologies function properly. Next we will introduce the visualization prototype that we implemented in order to present our methodologies to the end-user in a visualization application. Following the experimental study and the discussion on research question will be given and then we will talk about the threats to validity in

the end.

4.2.1 Proposed Visualization Methodologies

In this section, we will list the proposed visualization methodologies. First we will talk about interoperability in our data provenance visualization methodologies, second we will give the details of our user-centric and interactive visualization. Then we will explain our proposed visualization methodologies in following order, Provenance Data Summarization Visualization, Provenance Data Comparison Visualization, Provenance Data Dependency Querying Visualization and Streaming Provenance Data Visualization.

4.2.1.1 Interoperable Data Provenance Visualization

Here, we propose our suggested methodology for interoperable visualization and address the needs that we outlined in Section 1.4.2. To support interoperability as a quality feature, the provenance objects must be identified using a well-defined web semantic use. The W3C PROV-O ontology model is used for this study because it has a complete model specification, and similar research in application use domains is being used as reference. It complies with the data management standards in the authorized application domains that we listed in Section 1.3. Due to its expandable structure, the PROV-O Ontology provides a considerable level of interoperability for data interactions while also satisfying the requirements for provenance representation. Examples of these components include activity, agent, and entity. In Section 2.2, more information on the W3C PROV-O ontology model is provided.



Figure 4.7 PROV-O items' individual representation

To be able to design our conceptual model and to be able to implement the visualization approaches, the PROV-O ontology language is selected as the ontology of this study. Data preparation process is facilitated to import provenance data to our visualization tool in order to convert provenance data to PROV-O compatible nodes and edges as depicted in Figure 4.7. The details of the visualization framework is

given in Section 4.3.

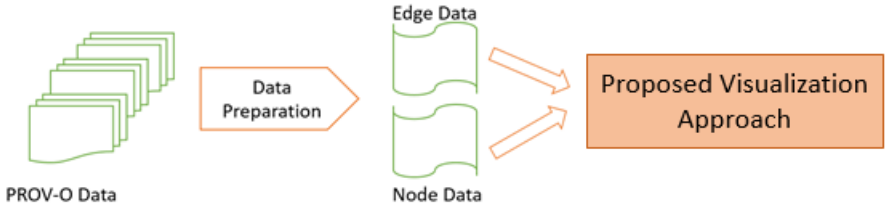


Figure 4.8 PROV-O compatible data importation

PROV-O data import functionality is implemented and proposed to the user inside our visualization toolbox. Node data includes the Agent, Entity and Activity ontology components while the edge data includes the relationships between components. Data preparation process converts provenance data files to the node and edge data to be imported to the visualization environment as depicted in Figure 4.8.

An example provenance data and the conversion samples are given in Figure 4.9. On right, there is a sample provenance data includes both provenance component definitions and the relationship definitions in a xml format. After the preparation process generated the node and the edge data to be visualized as shown in the middle including the node details as shown on the left.

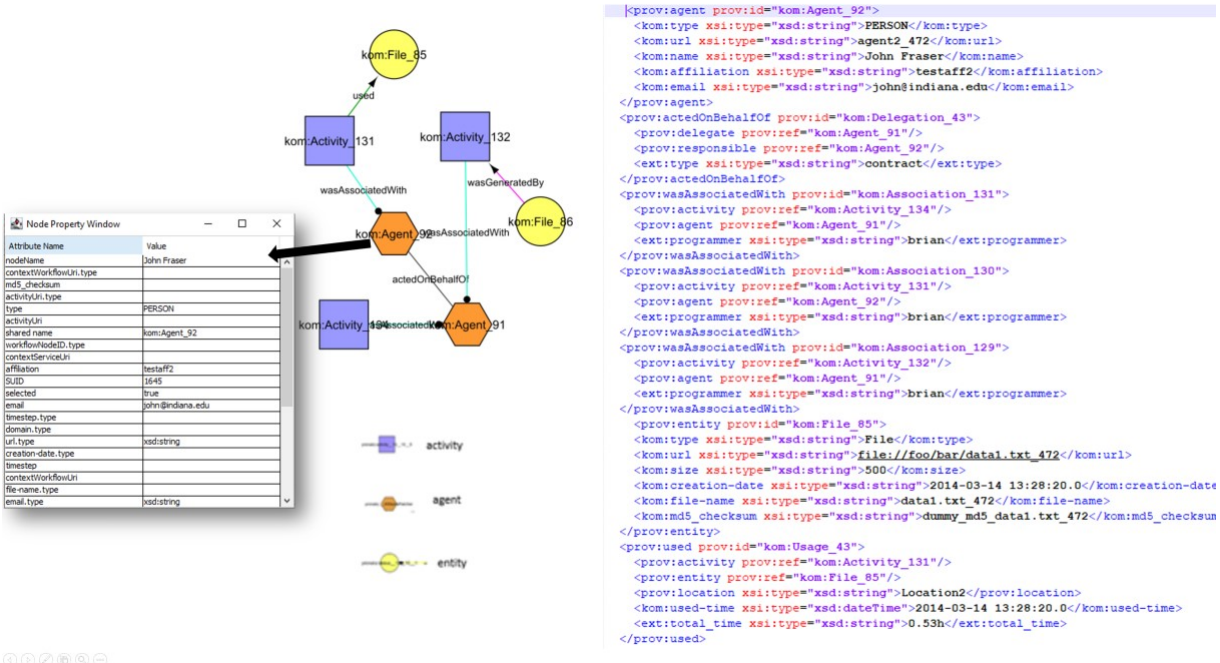


Figure 4.9 Data conversion sample

4.2.1.2 User-Centric and Interactive Provenance Visualization

As stated in Section 1.4.2 our main goal is to study on a set of visualization methodologies to enable end-users to better understand of provenance data. In order to achieve this goal, a visualization study should take user navigation models and HCI principles. The fundamental features are designed and implemented on an open source visualization framework to support basic user navigation under this subsection.

While building the HCI of the proposed system, we follow ten heuristics of Nielsen mentioned in Section 2.4 and we provide user interface solutions that employs all these heuristic principles. Now we will list the Nielsen's heuristics and our solution to meet them in the following.

To meet **Visibility Of System Status**, the visualization solution gives proper feedbacks stating the state of any user action such as provenance data importation status, comparison status or completion of dependency querying action in the information panel. **Metaphor Usage** is utilized in the user interface design properly. Especially in streaming visualization approach, the start, stop actions and forward backward movement in time actions require metaphore usage to speak with the users' language by having the real world icons on the user interface toolbox.

User Control and Freedom heuristic is applied for end-user not to lose control on visualization.For example when a provenance data is imported to the system, after some actions instead of importing the data to the system from scratch, the provenance data can be regenerated with an undo action. Each action can be triggered again and after executing the visual methodologies the data layout can be restored using the proper user interfaces. In the comparison methodology, all the relevant parameters and threshold values to be used in the comparison algorithm are opened to user for specification while the cluster count is a configurable value for the summarization algorithm to provide the freedom to the end-user.

In order to support **Consistency and Standards**, along with the user interfaces and functionality, the design and phrasing are all consistent. A naming convention is used to offer the user interface names. Unless the user changes their selection, an applied visualization style is constant across all user interface elements. The visualization tool reliably displays each ontology object independently as seen in Figure 4.7. For the sake of **Error Prevention** the mandatory actions are specified in detail by guiding the user to the next step, enabling/disabling user interfaces appropriately, and providing the user with appropriate feedback to avoid incorrect situations. The user cannot begin the comparison process without importing the two provenance graphs to be compared to the visualization environment. Additionally, while importing the data, the user is

informed of data inconsistency errors and is not allowed to proceed with the import activity.

To achieve **Recognition rather than Recall**, we concentrate on making objects, actions, and options clear, visible, and readily retrievable in the visualization environment in order to reduce user memory load. Instead of a single value entry, the value selections are offered with retrieved drop-down lists. For instance, drop-down values within acceptable bounds can be used to define node edge weight values and the threshold values in the advanced mode comparison. **Flexibility and Efficiency of use**, the idea of reducing steps for efficiency and flexibility is advocated. For instance, it only requires the user to choose a node from the visualization frame and one of the backward or forward options from the querying toolbox to run the backward, forward dependency querying. By permitting several node selections at once, the node property window can be launched at any moment and displays the details in a flexible manner as seen in Figure 4.10.

Aesthetic and Minimum Design is used while designing visual layouts while taking alignment, contrast, and brightness criteria into account. To accomplish the functional goals, the functional groupings are established in the user interface using fewer user interface components. **Help Users Recognize, Diagnose and Recover from Errors**, user comments and error messages are written in straightforward language that defines the issue clearly via warning boxes, information and error messages. **Help and Documentation** can be provided with a straightforward Help option that includes a text description of the system's functioning. The help should be simple to read, easy to comprehend, and easy to search for, with a focus on functionality.

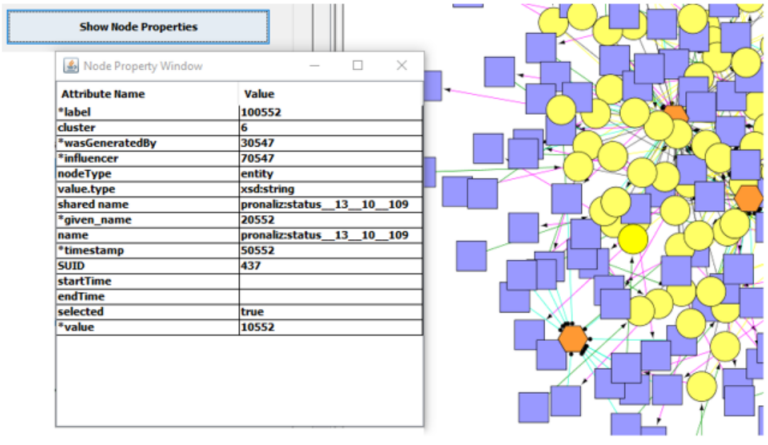


Figure 4.10 Example for node property window

4.2.1.3 Provenance Data Summarization Visualization

In this subsection, in order to address the summarization need in Section 1.4.2, we propose our summarization visualization methodology. A summary methodology is suggested to provide concise summaries and better comprehend their properties while considering large-scale provenance graphs that are PROV-O compliant. The amount of data is growing quickly, especially in the e-Science field. This data is noisy and has heterogeneous, dynamic, and distributed features. To create an effective and user-friendly prototype, e-Science provenance data is summarized in light of data expansion across the data life cycle. Figure 4.11 presents an illustration of a sample summary visualization.

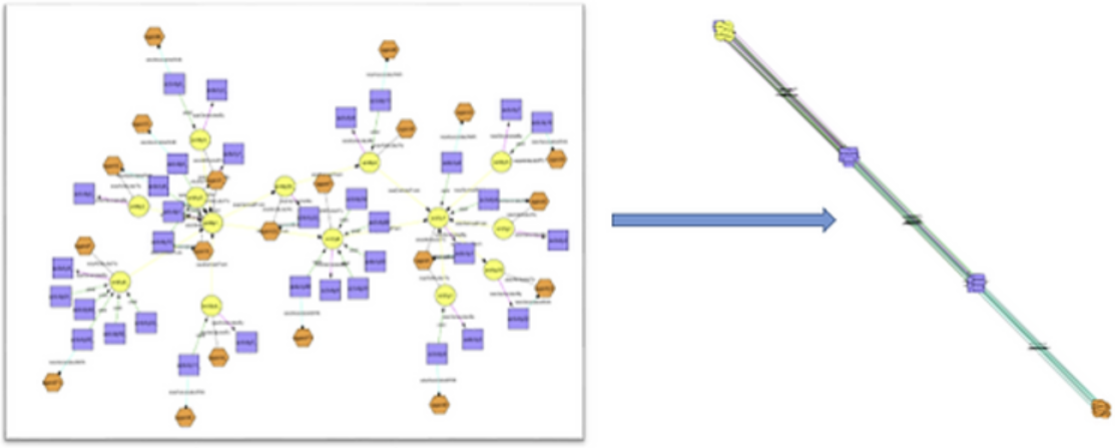


Figure 4.11 Provenance data summarization sample

In the unsupervised clustering, data sets are divided into different categories based on similarity, and a group is a collection of related data. To create a clear depiction of the commonalities between node properties, a property-based clustering is used in this study. There are several approaches to clustering exists. The three common clustering approaches are given in the following.

Centroid Based Clustering: Centroid-based clustering groups the data into non-hierarchical clusters.The data is organized into non-hierarchical groupings using centroid-based clustering[79]. The aspect of determining the distance between the objects in the data set under consideration is the key underlying theme in all centroid-based methods. Typically, one of the Manhattan, Minkowski, or Euclidean distance measuring processes is used to obtain the fundamental component of distance measurement. A representative of the approach is the K-Means algorithm in this study.

Even for large amounts of data, the K-Means algorithm is recognized as a popular clustering method that divides data into a predetermined number of clusters. K-Means

is scalable due to its iterative form, and in their paper [79], a novel parallel K-Means approach based on a straightforward computational model was presented.

Hierarchical Based Clustering: The fundamental tenet of hierarchical clustering, sometimes referred to as connectivity-based clustering, is that items are more closely associated with one another than with other things that are farther away. Based on the distance between "items," these algorithms link them to create "clusters." The greatest distance required to connect cluster members can be used to describe a cluster in a significant part. The term "hierarchical clustering" refers to techniques that offer an extended hierarchy of clusters that merge with one another at specific distances rather than a single division of the data set. Different clusters will develop at different distances. As representative of the approach is Bisecting K-means algorithm in this study. The Bisecting K-Means algorithm which is an iterative of K-Means is considered as a type of top-down hierarchical clustering, in which all clustering begins in a single cluster and is repeated recursively for each split. Until the desired number of clusters is obtained, a cluster is repeatedly divided into two new clusters. As an estimator, Bisecting K-means develops a common model called the Bisecting K-Means model. [80] [81].

Distribution Based Clustering: This clustering technique[82] assumes that data are composed of Gaussian-like distributions. Three Gaussian distributions are used by the distribution-based approach to group the data. As one gets away from the distribution's center, the likelihood that a point belongs to the distribution decreases. The bars represent these diminishing probabilities. If you don't know what kind of distribution your data have, try a different strategy.

Data points from one of k different Gaussian subgroups are combined to create a composite distribution known as a Gaussian Mixture Model. The maximum-likelihood model can be produced via the expectation-maximization technique. As an estimator, GaussianMixture develops a GaussianMixtureModel that serves as the base model. Given a set of data, the expectation-maximization (EM) algorithm is used to infer the maximum-likelihood model. The maximum likelihood and maximum posteriori estimates of the statistical model parameters are found iteratively using the EM approach, where the model is built on hidden, unobserved features.

In order to explore the workflow data provenance and learn about the time-based properties of the provenance data, the "occurrence time" property, as defined in the PROV-O ontology model to cluster nodes, is utilized for the summary. The user's primary area of interest, in this case, is provenance presence, therefore an activity item's data-centered "creation time" aspects are the ones that are highlighted. Figure

4.12 provides a pseudo-code for the PROV-O summarization algorithm.

```
1: get all activities in provenance data
2: sort activities in ascending order
3: get data entities of activities in ascending order
3:  define k-means k value
4: generate k group from data entities
5: for all k group
6:    for all data entities
7:          set data entity cluster
8:          set data entity adjacent agent nodes
9:          set data entity adjacent activity nodes
10:          manage overlapping for visualization
11:    end for
12: end for
```

Figure 4.12 Pseudo code for provenance summarization algorithm

In order to store and process the growing amount of data in recent years, new frameworks are required. We look for potential ways to guarantee scalability for the summarization as provided in 4.2.4.1 in order to increase our visualization capabilities and performance for large amounts of data.

4.2.1.4 Provenance Data Comparison Visualization

Comparing provenance graphs, as required by Section 1.4.2, is one of the primary objectives of a data provenance inquiry. It is crucial to compare the two graphs, especially in the field of e-Science, in order to find anomalies and commonalities. In this paper, an approach is proposed that broadens Chen's graphical comparison algorithm[40]. They compare two OPM provenance trees using the Direct Node Classification (DCA) algorithm in their research. By comparing graph nodes, Isomorphisms between graphs and subgraphs can be found through direct classification of node attendance. The process first assessed the data indicating the probability of a node's predicted participant in another graph. The evidence is based on measures that are specific to each node, such as node connectivity and node and edge attributes. The W3C Open Provenance Model, an application-specific model, is coupled with a methodical specification. An improved DCA algorithm is proposed in the current study which is compatible with the W3C PROV-O data model. We use provenance documents that are PROV-O compatible, hence the method is known as DCA-Prov, while Chen's version is designated as DCA-Chen. The following provides information on the DCA-PROV algorithm in depth.

Figure 4.13 displays a sample of comparison visualizations. The comparison above is the sample of the DCA_{label} mode comparison. The comparison down exhibits the DCA_{props} comparison with the original graphs on the left and the comparison results on the right. Following the comparison, the green nodes represent similar nodes, while

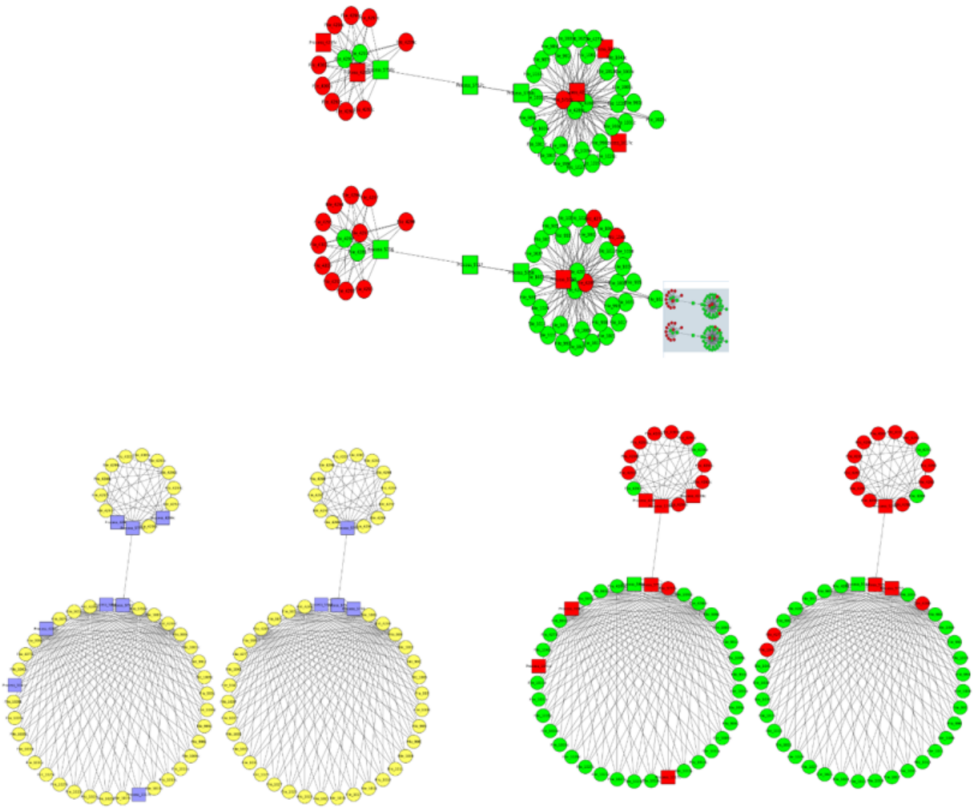


Figure 4.13 DCA-Prov provenance data comparison

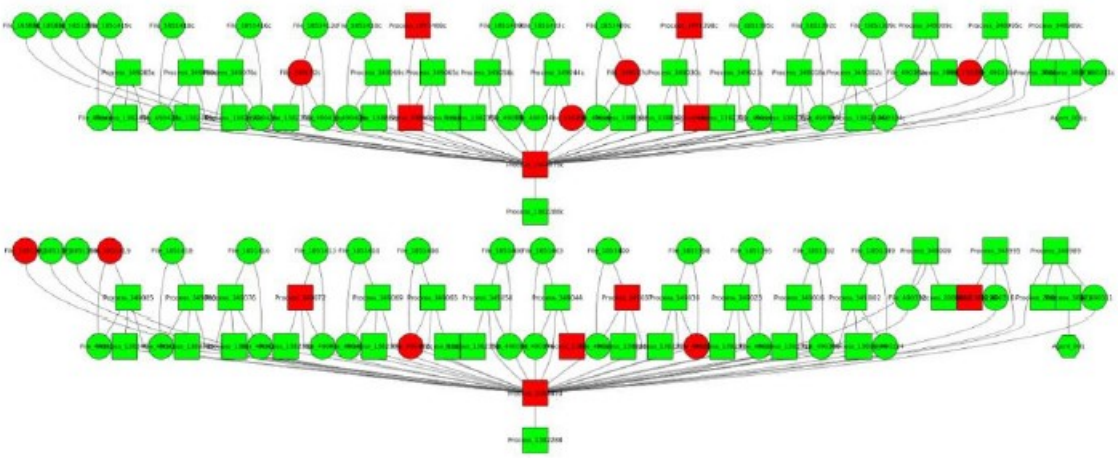


Figure 4.14 DCA_{label} provenance data comparison

Table 4.2 Differences and similarities of DCA algorithms

	DCA-Original	**DCA-Chen**	**DCA-Prov**
Symmetric Edge Iteration	No	Yes	Yes
Node Item Type Comparison	No	No	Yes
Minimum Threshold Usage	No	No	Yes
Configurable Minimum/Similarity Threshold	No	No	Yes
Configurable Node/Edge Property Weights	No	No	Yes

the red nodes represent distinct nodes. Figure 4-14 depicts the DCA_{label} comparison on 10GB Noisy Semi-Synthetic Dataset.

Enhanced DCA Algorithm (DCA-Prov) The direct classification of nodes verifies the isomorphism of subgraphs and graphs. The method starts by evaluating the evidence that indicates the likelihood of a predicted node's presence in another graph. The calculation of the evidence depends on node-specific information such as edge and node properties and node connection. The procedure then searches for a node-to-node mapping and rearranges nodes to enable direct comparison of the resulting graphs. Under different conditions, the method uses polynomial order to produce approximate solutions for subgraph isomorphism with at least 95 percent performance.

To solve the isomorphism issue between subgraphs and graphs, the DCA method proposes the calculation of attendance rates. The original DCA technique is enhanced by Chen et al. [40] in their paper by using the greedy approach to capture the node attendance when creating a subgraph and including symmetric node iteration regardless of the size of the graph. The graph comparison algorithm is investigated in this study referencing from Chen et al's approach as a foundation and by contributing to the improvements, which are listed in Table 4.2. There are node attendance calculation and node attendance check phases in the proposed PROV-O graph comparison algorithm. The attendance value between node pairs by having one node from each compared graph for node-to-node comparison is calculated in the first step. The second phase verifies the highest possible attendance across all potential node pairs.

The DCA algorithms' comparison is provided in Table 4.2. In order to compare all edge values of the compared node value, the edge comparison's loop size should be set to the higher edge number between two nodes. This is because DCA-Chen and DCA-Prov both use symmetric edge iteration. To provide a cost-effective solution, we included the node type comparison to the comparison stages and a minimum threshold value to determine minimum similarity. Additionally, the user may correctly configure the

comparison threshold values. The following algorithm explanations will provide the details.

Node Attendance Calculation: The potential of a node's attendance in the other network is described by the attendance value, that can be thought like a similarity score between one node pair. An attendance score for a node is given in a scale of 0.1 to 1.0. A closer score of 1.0 signifies greater similarity between node pairings. The proposed algorithm below illustrates pseudo-code of the revised graph comparison algorithm. It is known as DCA-Prov since it is enhanced for PROV-O-compatible provenance documents. According to the pseudo-code in Figure 4.15, DCA-Prov includes two distinct modes, DCA_{label} and DCA_{props}.

Proposed Algorithm: The attendance calculation outlined just before is the basic foundation of our proposed methodology. The provenance graphs' node pair and edge attributes are used to identify the similarity.
Assumption 1: The parameter of the DCA-Prov algorithm's execution mode is either DCA_{label} or DCA_{props}.Lines 3–9 show that, similar to Chen et al.'s approach, regardless of the mode, DCA_{label} or DCA_{props}, we use a symmetric approach to iterate the edges of the node pair and assign the greater edge count value to N1 variable.

In contrast to DCA_{label}, which only takes into account the node type and related edge type attributes, DCA_{props} takes into account all edge and node properties. DCA_{label} is free of attendance calculation because we simply compare important parameters to determine if the node pair is different or similar. Lines 20–26 and 37–39 depicts that after comparing node/edge type in DCA_{label} mode, we explicitly designate node pairs as similar or not similar. For instance, we identify two node pairs as not comparable if they include "agent" and "entity" node types or "wasGeneratedBy" and "wasDerivedFrom" edge types.

Node and edge attributes are taken into account in DCA_{props} mode in detail by considering the weight set to check two workflows.In line 11, it is demonstrated how to do this. Since the provenance node type (agent, activity, and entity) has the primary factor, we first examine the node type similar to DCA_{label} mode. We contend that skipping the comparison and not computing the attendance reduces the comparison complexity when the node types of the compared node pairs are not the same. The cost of comparison for the attendance list is n*m when there are two graphs of sizes n and m. The comparison count increases to n^2.

The size of provenance items is shown as x: entity count, y: agent count, and z: activity count in Theorem 4.1. The cost is $(x + y + z)^2$ when we assume that the counts of the

```
Algorithm: DCA_E graph comparison
    Input: Two provenance graphs G1 = (V, E) and G2 = (V, E)
which contain node and edge items to compare similarity.
    Output: Similarity mark for each node {similar, not similar}.
    1 For each Node1 in Graph G1 do
    2   For each Node2 in Graph G2 do
    3    if (Node1 incident edge# > Node2 incident edge#)
    4            N1 = Node1 incident edge #
    5            N2 = Node2 incident edge #
    6       else
    7            N1 = Node2 incident edge #
    8            N2 = Node1 incident edge #
    9        end if
    10       init attendance (Node1,Node2) to default value
    11       if (Node1.type == Node2.type)
    12         For each Edge1 incident to N1 do
    13           For each Edge2 incident to N2 do
    14             If DCA_props?
    15              Compare connectivity along Edge1, Edge2
    16              Compare edge props of Edge1, Edge2
    17              Compare adjacent nodes props of Edge1, Edge2
    18              Update attendance
    19             End If
    20            If DCA_label ? and
    21             Edge1.type !=  Edge2.type  OR
    22             Edge1.all adjacent nodes.type !=
                              Edge2.all adjacent nodes.type
    23                Mark Node1 and Node2 NOT SIMILAR
    24             Else
    25                Mark Node1 and Node2 SIMILAR
    26              End If
    27     End for (13)
    28      Keep matched edges and adjacent nodes comparison
    29     End for (12)
    30      If DCA_props?
    31        Compare node properties of Node1, Node2
    32        Compare node edge count of Node1,Node2
    33        Update attendance
    34        If attendance above MINIMUM_THRESHOLD
    35            Update attendance_list for Node1,Node2
    36        End If
    37      Else If DCA_label ?
    38        Mark Node1 and Node2 NOT SIMILAR
    39      End if
    40 End For
    39 End For
    40 Node Attendance Check ()
```

Figure 4.15 Enhanced-DCA Node Attendance Calculation Pseudo code

provenance elements are equal.

The calculation demonstrates that the cost of attendance calculation and cost of comparison are decreased by just taking into account the similar nodes to $(x)^2+(y)^2+(z)^2$.

Theorem 4.1. *Comparison complexity*

$$(x+y+z)^2 = x^2+y^2+z^2+2(xy+xz+yz) \tag{4.1}$$

DCA_{props} utilizes attendance-based comparisons. The attendance value is set to 1.0 initially, which perfectly matches the attendance value, in step 10 of the process. Using weight coefficients, step 18 updates the attendance value in accordance with the comparison properties as shown in lines 15–17.

In each comparison step, the attendance n_i , n_j is recalculated seen in Theorem 4.2. According to their respective weights, node properties (line 31), adjacent node properties (line 17), and edge properties (line 16) have an impact on attendance value in comparison.

Theorem 4.2. *Attendance value calculation*

$$\textit{attendance-value}_{updated} = attendance - value_{actual} * ((weight_{maximum} - weight_{different-property})/weight_{maximum}) \tag{4.2}$$

If the properties of node pairs' differ, as shown in Equation 4.2, steps at the 16th, 17th, or 31st lines update the attendance value removing the property weight from the current value of attendance. In order to eliminate computational cost for the attendance list calculations and comparison stages, a minimum threshold condition is added on line 34. Additionally, a minimum threshold which is a user-configurable value is employed in the node attendance calculation phase to establish the limit for attendance to be used for elimination in categorization. This reduces system complexity by defining a threshold in the node attendance check for similarity. In our visualization framework, a default value of 0.5 is given to it. Our experiment demonstrates that by selecting a 0.5 minimum criterion, we enhance the complexity of the similarity comparison at almost 60 percent.

Node Attendance Control: This stage is just the execution of DCA_{props} mode. The maximum attendance is gathered utilizing attendance combinations in the DCA algorithm.(n_{G1}, n_{G1}). Starting from the node pair has the maximum attendance, a greedy technique is used to check the neighboring value nodes against a predetermined threshold value. The proposed version lists the DCA-Prov node attendance combinations in descending order. A node match with the highest attendance value is designated as being the first. Only unmarked nodes are taken into account for attendance threshold control. The similarity criterion is established based on the results of our experiments and is set at 0.7 when two provenance data sets are compared. In the DCA_{props} approach, in our provenance visualization prototype, the user can set the threshold value.

By removing Chen's greedy method, we make a contribution to the DCA algorithm by taking into account all node pairs in two workflows. In order to underline the need for similarity, we design two DCA techniques named DCA_{label} and DCA_{props}. In DCA_{props}, the comparison is customized using user-defined property weight factors for each node and edge property. We reduce computational complexity and eliminate redundancy costs by having the node types at the start of the comparison and then by utilizing a minimum threshold value. Each decision value that is used for the algorithm, similarity threshold and minimum threshold can be defined by the user.

4.2.1.5 Provenance Data Dependency Querying Visualization

Here, we address the needs identified in Section 1.4.2 and propose a framework for provenance data dependence visualization. In order to establish a cause-and-effect link between PROV-O nodes and their connections, additional analysis and study on PROV-O data are necessary. In this study, Backward Dependency Querying (BDM) and Forward Dependency Querying (FDM) are proposed for PROV-O activity, agent, and entity elements in a workflow process.

Provenance Backward Dependency Querying : Information regarding the input components which the output depends on is known as backward provenance dependence. Finding data portion or input summaries matching a certain area of output can be made easier. Backward provenance dependence includes input and intermediate entities of agent, activity, entity items connected to output data production since the PROV-O model is used in this study. Scientific experiments, particularly e-Science procedures, produce large size provenance data over the course of hours or days of operation.

Backward provenance helps users troubleshoot issues and comprehend how a

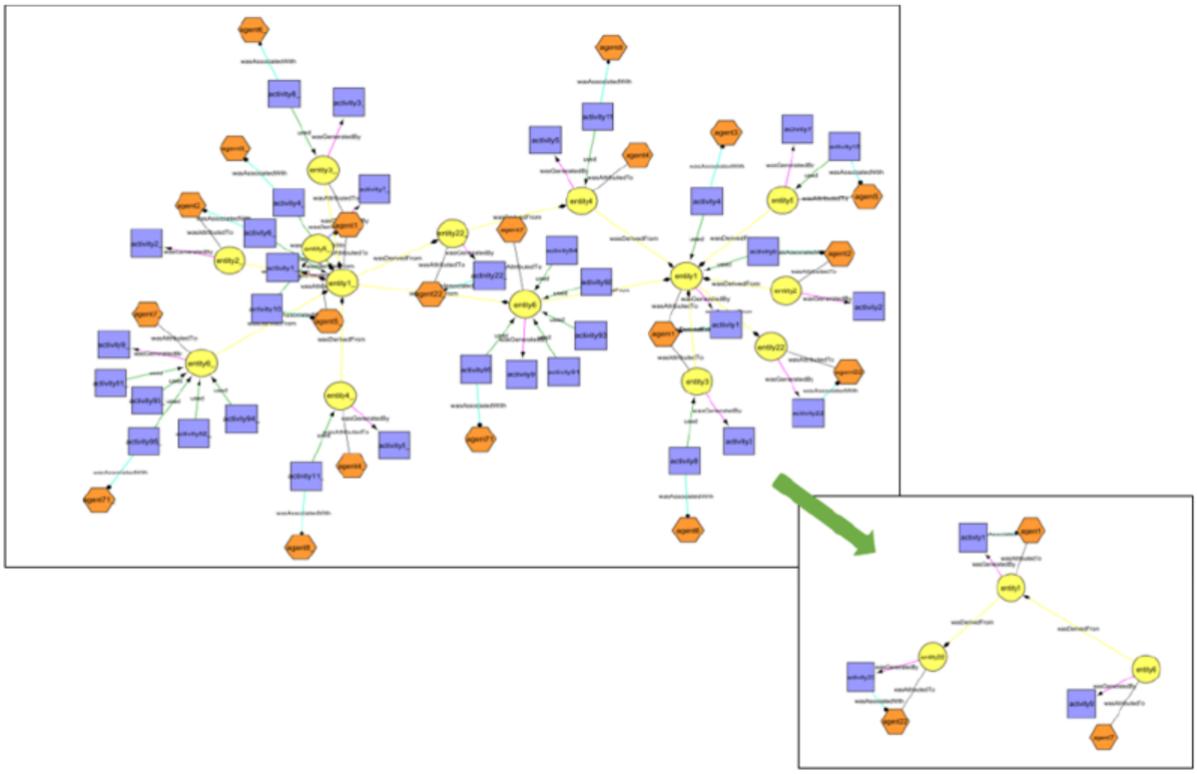


Figure 4.16 Backward Dependency Query Sample

particular outcome came about. It is also appropriate to use relationships of the parameters and input, intermediate, and output data to identify process failure steps. Backward provenance querying, for instance, is to be used for identifying the underlying causes of any unexpected entity state or activity that researchers detect while keeping track of the workflow execution. Based on Chen's BDM[40] technique, which handles the provenance data constantly as a ordered data stream, we provide an improved BDM algorithm. To answer the backward data requests of recent provenance nodes in the system, Chen maintains a dependency matrix. Using input/global parameters, the dynamic matrix is utilized to record and calculate the nodes' dependencies. A sample visualization on backward dependency querying is depicted in Figure 4.16.

Proposed Algorithm: Our BDM technique, which was covered in the preceding part, is the foundation of our proposed methodology. We update our backward dependency using the coming node and edge data as well as relationships.

Assumption 1: When new edge and node information is received to define source $node_i$ and destination $node_j$, and e $= v_i, v_j$, we presume that a backward dependence of node j already exists in our dependency matrix based on the temporal data order of provenance data.

Pseudo Code: Backward dependency update state

Each node is given a distinct node identifier to attach dependency information to a node, as shown in the pseudo-code in Figure 4.17, and BDM queries are carried out by running get backward provenance operations repeatedly. The dynamic dependency structure known as currentBDM maintains dependencies for each node id.

```
        Algorithm: Backward dependency update state
Input: New edge and node information that defines a relation
(element).
Output: Backward dependency matrix.
1   if element is an EDGE information
2     sourceNode <-- element.source
3     destinationNode <-- element.destination
4      if sourceNode identifier has changed?
5        remove all dependencies of old node id from currentBDM
6         cache all dependencies of old node id to cacheBDM
7     end if
8     update sourceNode identifier in currentBDM
9     if destinationNode is in application
10      add new dependency sourceNode id -->destinationNode id
                                                  in currentBDM
11    end if
12    if sourceNode id equal to destinationNode id
13      get backward provenance from cacheBDM to deps
14    else
15      get backward provenance from currentBDM to deps
16    end if
17    for each dependency in deps
18          add dependency to SourceNode in currentBDM
19    end for
20          end if
21 end function
```

Figure 4.17 Backward dependency querying pseudo code

Data structure size is decreased from "*node count*node count*" in the dependency dynamic matrix to "*node count * dependent node count*" in the key-value dictionary. O(N) is the complexity of the given algorithm. While maintaining the dependency information in a dynamic matrix struct in the BDM approach of Chen et al's, a dynamic dictionary struct is utilized in our modified BDM technique to reducing cyclomatic complexity and execution step number. The improved technique avoids the expense of adding a row for each column sequentially by adding an empty array to the dynamic matrix. Instead we utilize a dictionary struct in the situation that the sourceNode identifier is not in currentBDM as shown in line-8. In addition, we merely add a new key–value pair to the dictionary struct in O(1) time. A single assignment adds the dependency to the sourceNode id rather than iterating in line-17 and line-19 by adding an existing destination (deps) dependency to the source node.Table 4.3 includes a comparison of the enhanced BDM (E_BDM) and the Chen's BDM (C_BDM) approaches.

Table 4.3 Comparison of BDM Algorithms

	C_BDM	**E_BDM**
Update BDM with incoming node/edge info	Yes	Yes
Assuming each node id for dependency is in the BDM	Yes	Yes
Archiving an earlier source Dependencies on nodes	Yes	Yes
BDM dynamic dependency data size	$nodecount^2$	$nodecount * dependentnodecount$
Adding a new sourceNode identifier to BDM	Costly	Cost-effective
Adding a new dependency complexity	O(N)	O(1)
Recursive dependency	N/A	Yes

Provenance Forward Dependency Querying : In addition to the introduced backward provenance dependency algorithm, a brand-new querying is also presented for the PROV-O forward dependency to facilitate dependency analysis. Intermediate and output nodes analysis based on a chosen input node and the subsequent history of this particular node are both examples of forward provenance dependence. Agent, activity, and entity components make up the input, intermediate, and output nodes in our PROV-O model when querying the future. To debug and analyze large-scale workflows, e-Science researchers use querying techniques to help them understand the possible effects of a PROV-O item in the data lineage. To analyze the results of erroneous PROV-O items and their effects on a process or workflow, the researcher is enabled to utilize the suggested forward dependency querying. The suggested FDM approach is based on the improved BDM methodology, maintaining a dependency data matrix that iterates the data nodes to maintain the relationships with a dynamic manner by use of a dictionary.

4.2.1.6 Streaming Provenance Data Visualization

Here, we present our recommended methodology for visualizing the provenance of streaming data and responding to the issues we identified in Section 1.4.2. The publish-subscribe messaging system receives streaming notifications in the form of selected ontology model from various distributed environments to enable distributed processing of streaming provenance data. To cache the visualization data for presentation in our visualization environment, the provenance data elements are gathered and stored in in-memory storage as illustrated in Figure 4.18.

4.2.2 Datasets

We introduce the datasets we used ih this thesis, a massive 10GB semi-synthetic provenance database [83], a social workflow dataset and a NASA AMSR-E provenance archive dataset [26].

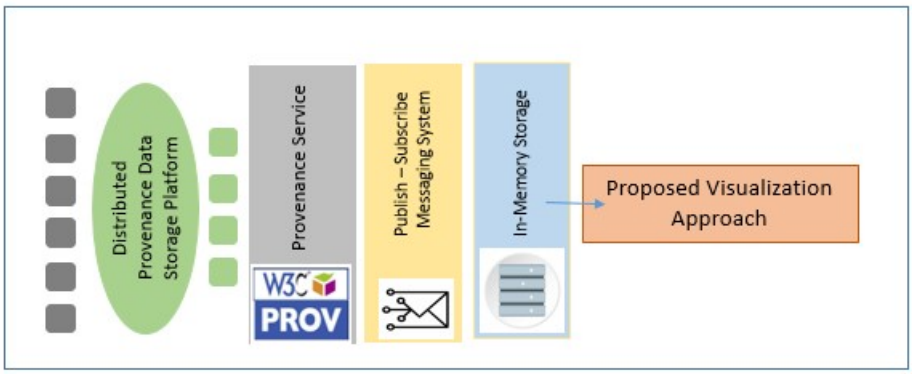


Figure 4.18 Provenance stream visualization solution

4.2.2.1 10GB Noisy Provenance Database

10 GB Noisy Dataset: A semi-synthetic dataset created by Indiana University using real world workflows using WORKEM emulator to generate 10 GB provenance database. Almost 48000 workflow execution are used by modeling 6 real workflows with erroneous patterns. LEAD North American Mesoscale (LEAD NAM)(weather), SCOOP ADCIRC (coastline), NCFS (ocean), Gene2Life (bio), Animation (CS), and MotifNetwork (bio) are the workflows that exist in a 10GB noisy database [83]. To export these data to OPM-compliant provenance data, the Karma provenance tool was used and 2.1 GB of data is used as OPM graphs.

In our study, since we address W3C PROV-O compatible provenance system, provenance data generation from OPM to PROV-O is handled in our data conversion phase. As a result, we produce 1.2 GB of semi-synthetic PROV-O XML data based on a 10 GB Noisy dataset.

4.2.2.2 Social Provenance Dataset

A provenance database must satisfy a number of large-scale, diversity, and realism requirements, according to [84]. To enable large-scale research, a provenance database should contain a significant number of provenance records. The provenance database should be compiled from a variety of procedures with various sizes, breadths, and lengths. The workflows assembled to produce the provenance should exhibit failure characteristics. We include usability as an additional need in addition to the ones already mentioned. We contend that a provenance database should fulfill both its domain-dependent criteria as well as its general requirements for a provenance database.

In this work, a generated social provenance database in study [85] is used that complies with the criteria listed above. By creating three separate forms of social

provenance, each of which represents a different scale of social interactions, we were able to satisfy the diversity criteria. 100, 1K, and 5K are the different social contact categories. We generated 100 social workflow execution traces, one for each type of social contact. By producing the identical dataset with a 10 percent notification failure and 10 percentage execution failure rate, we satisfied the realistic criterion. For scientific datasets, Cheah et al. created a noisy, 10 GB provenance database with failure characteristics [84].

4.2.2.3 NASA AMSR-E Provenance Dataset

Provenance data generated from NASA National Snow and Ice Center Distributed Archive Center daily, weekly and monthly based on NASA's Aqua satellite is used to produce OPM provenance data from Indiana University [26]. These are several daily, weekly and monthly workflows captured to generate large scale graph data. Daily Rain graph is regarded as the smallest provenance graph with the size of 11 KB, while the largest graph called Montly Rain graph's size is 9 MB. Similar to 10 GB Noisy dataset, we realize a data generation operation to generate PROV-O compatible dataset from Indiana University's data archive in OPM ontology model.

4.2.3 Prototype Implementation

A platform for molecular interaction networks and biological pathways to be seen, integrated with annotations, gene expression profiles, and other state data is called Cytoscape. Even though Cytoscape was first created for biological study, it today serves as a universal platform for sophisticated network analysis and visualization. One of Cytoscape's strongest points is its ability to use plug-ins to develop new visualization techniques.

An experimental prototype is developed in this work for assuring visualization capabilities , for examining the applicability and outcomes of the suggested methodologies. To achieve a user-friendly, interactive, and to manipulate provenance data display by taking into account user and HCI concepts, common visualization components are utilized. We implemented our visualization methodologies on top of Cytoscape which is labeled as Proposed Visualizaton Approach in Figure 4.18. As shown in Figure 4.19, the prototype was created using the Cytoscape visualization framework and includes standard functionality for data integration and display. The reason for using Cytoscape tool is an extendable visualization open source framework and we are able to implement our methodologies as plug-ins.

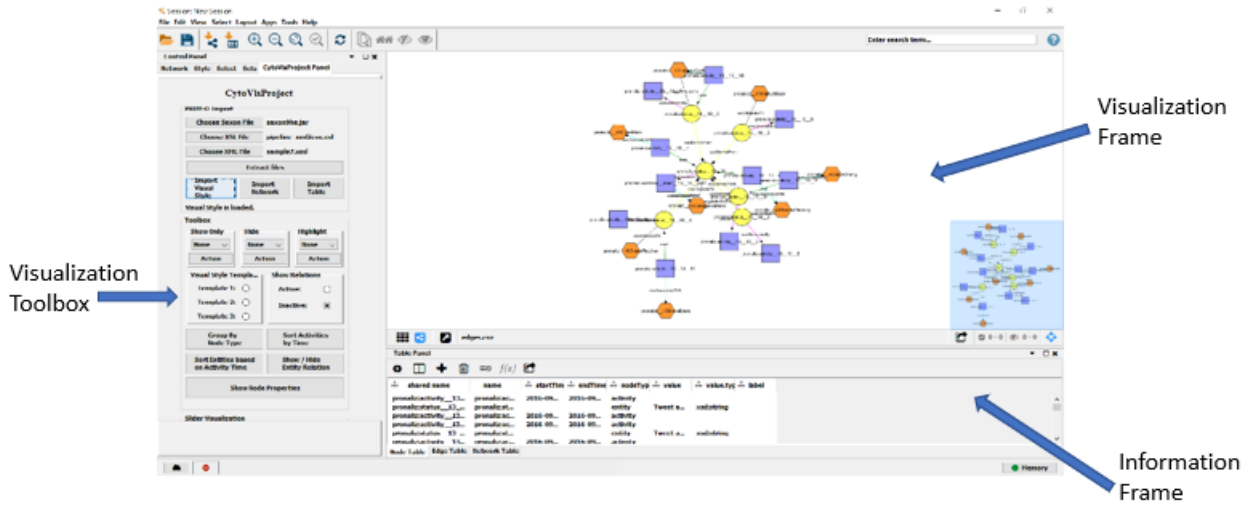


Figure 4.19 Visualization Prototype

The left toolbox contains the individual panel for each visualization style as well as the accompanying user interface elements. For instance, the user is able provide the required information by utilizing the graph comparison user interface, including the configurations for the node and edge weights as well as the DCA_{label} or DCA_{props} mode. The prototype includes every option so that the user would be able to choose for visualization methodologies. This further integrates the user with the system as much as is practical. The usability of the suggested visualization tool is assessed as discussed in Chapter 5 by facilitating a thorough usability study.

The codebase of the project **[86]**, the research page **[87]**, and the demo video **[88]** are prepared accordingly. We implemented our prototype in Java language version 11.0.7 using Cytoscape API version 3.9.1. Our detailed environment set-up instructions can be found on the research website above.

4.2.3.1 Scalability Study on Provenance Data Summarization

New processing frameworks, particularly for scalable platforms for future expansion, were required for the high-scale data processing requirements. The following presents the results of the study on scalable clustering techniques. The framework for performing adequate distributed computing for Big Data utilizing in-memory concepts is Apache Spark. New processing frameworks, particularly for scalable platforms for future expansion, were required for the high-scale data processing requirements [89] [90]. The framework for performing adequate distributed computing for Big Data utilizing in-memory concepts is Apache Spark.

A quick and well designed platform, using novel ML algorithms, can make it

easier for users to load data into memory and analyze it frequently online. The drawbacks of earlier approaches like Map Reduce/Hadoop, like the restrictions on linear data and disk use, were taken into account when developing Apache Spark. The Apache Spark clustering environment is depicted in Figure 4.20. The purpose of MLlib, Sparks' machine learning library, is to make machine learning scalable and applicable by offering ML methods, features, and utilities. To evaluate the scalability on graph summarization, the three novel clustering techniques listed below—K-Means Clustering, Bisecting K-Means Clustering, and Gaussian Mixture Model Clustering—were used.

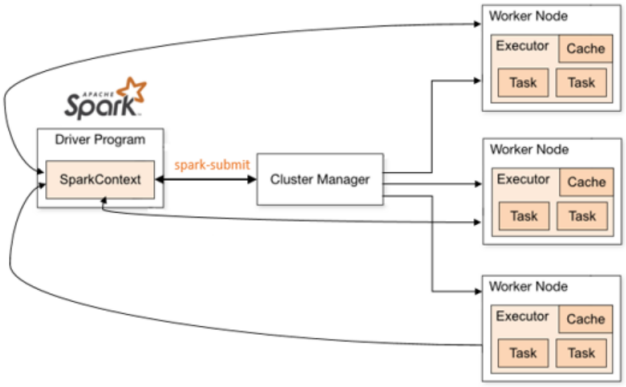


Figure 4.20 Spark clustering environment

Number of Clusters: The optimum cluster number for the data set in question must be determined while partitioning data sets, which need the user defined cluster number as k value. In the literature, there are some definitions on the ideal cluster number, although they are frequently combined with other evaluation methods rather than being chosen over others. The Elbow approach is utilized in this study to find an elbow in the function of clustering cost by repeatedly running the given clustering algorithm while experimenting with different selections of k for the data fractions [91].

4.2.3.2 Streaming Provenance Data Visualization

While implementing the proposed streaming visualization methodology in Section Here, we propose the implementation of our suggested methodology for streaming data provenance identified in Section 4.2.1.6. The technologies are selected to fulfill the streaming requirements in the proposed solution. As a publish-subscribe messaging system, the Kafka message broker system is utilized in order to have online streaming notifications. Kafka tool is a distributed message system with high throughput that may be utilized similarly to a general message broker system. Large amounts of data can be accepted and stored there regardless of the users.The provenance data messages are collected by a Redis memory cache to visualize data in

our visualization environment.

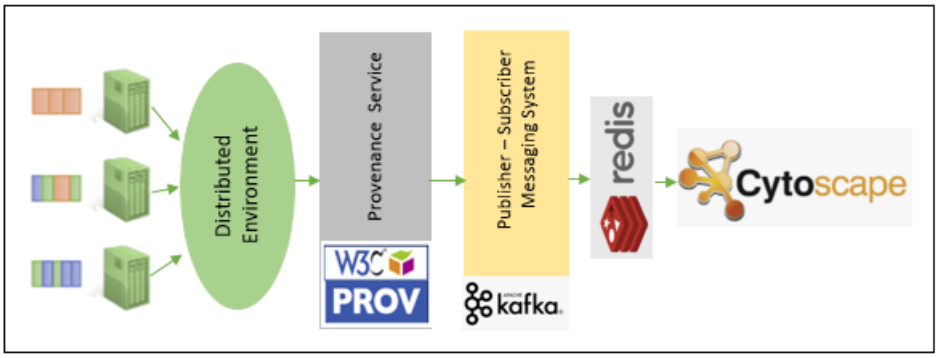


Figure 4.21 Provenance visualization streaming flow

The Figure 4.21 illustrates the technologies in order to meet the requirements shown in Figure 4.18. The individual Kafka client's distribution generates provenance data as Publishers and the main Kafka subscriber gets these provenance data concurrently in the system. Messaging topics are defined to communicate with publishers and subscribers according to well-known publisher-subscriber patterns. The provenance data messages are collected by Redis memory cache to visualize data in our visualization environment Redis is an in-memory remote database that offers high performance, replication, and a unique data model to produce a platform for solving notification problems.

4.2.4 Experimental Study Results

In this section, the experimental results for the proposed visualization methodologies are presented in the following. First the data summarization scalability experimental results will be given and then the provenance comparison and dependency querying results will be presented.

4.2.4.1 Experimental Study for Provenance Data Summarization

Apache Spark framework MLlib libraries are used to perform distributed scalability assessment on provenance data summarization using three novel algorithms, K-Means Clustering, Bisecting K-Means Clustering and Gaussian Mixture Model Clustering. In addition to the K-Means clustering, K-Means with RDD mode is also issue of scalability experiments. The Spark MLlib K-Means having RDD implementations is called paralleled variant of a K-Means++ method and scalable K-Means. The details on the assessment such as data, environment, and configurations, and the experimental results are given in the following.

Table 4.4 Scalability Data Sets

Data Size	PROV-O Node Count
1M	6000
10M	55000
50M	255000
100M	509000

Environment: For scalability, a Big Data, Spark cluster environment was built up with 3 worker-nodes, as illustrated in Figure 5. The resources have an Intel® Core TM i5-10210U CPU running at 2.10 GHz and 8 GB of RAM. Each worker-node is set up with two cores and 1024 KB of RAM.

Data: The Table 4.4 below shows the node counts created for the visualization sample data for the scalability research and each PROV-O item has ten features as a feature set.

Number of Clusters: The Elbow method is used to find optimal cluster count, which is a common issue in summarization to specify k numbers. After executing multiple iterations of the K-Means algorithm, the number of cluster, k is established by experimenting with various k values for the data slices and searching for the elbow in the clustering cost method [91].

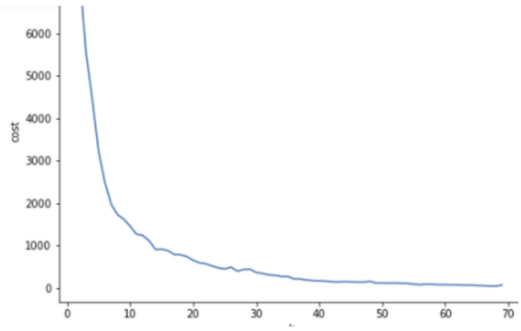


Figure 4.22 Elbow function results for optimum cluster number

For the dataset illustrated in Table 4.4, the elbow technique is used to determine the optimum cluster count, and as seen in Figure 4.22, processing the entire data set with k=50 yields very little change after k=50. Scalability assessment configurations are given in Table 4.5 including clustering parameters such as the iteration count, distance methodology and the library information.Each algorithm is tested ten different times, and measurements are kept to obtain the mean value that will be used in the scalability study. The experimental results are depicted in the following figures showing the clustering time in seconds for each dataset in Table 4.4 individually.

Table 4.5 Clustering Algorithms and parameters for scalability study

Clustering Name	Iteration	Distance	Library
KMeans	1000	euclidean	Apache
KMeans with RDD		euclidean	Spark
Bisecting KMeans		euclidean	In
Gaussian Mixture		N/A	Python

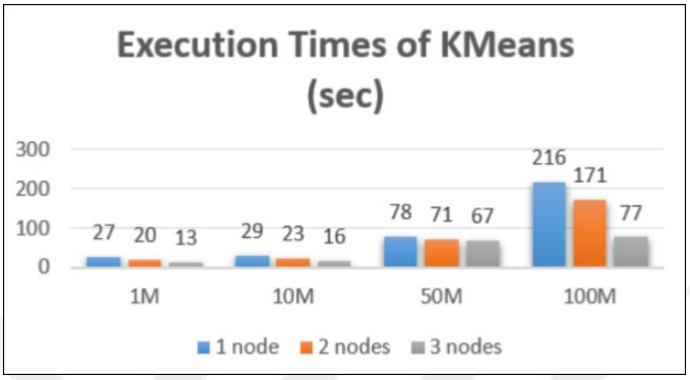


Figure 4.23 pySpark KMeans Clustering scalability results

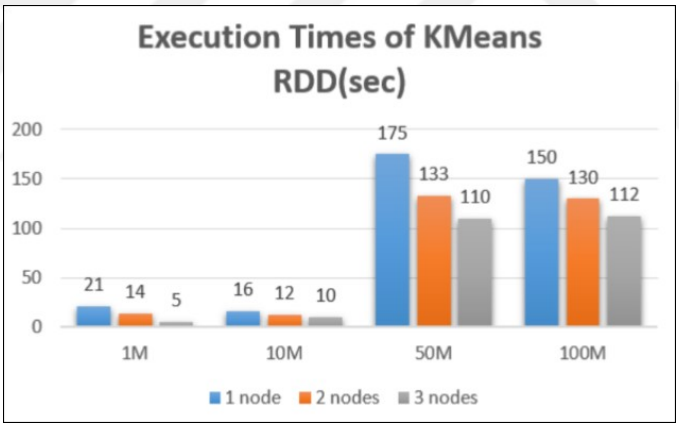


Figure 4.24 pySpark KMeans-RDD clustering scalability results

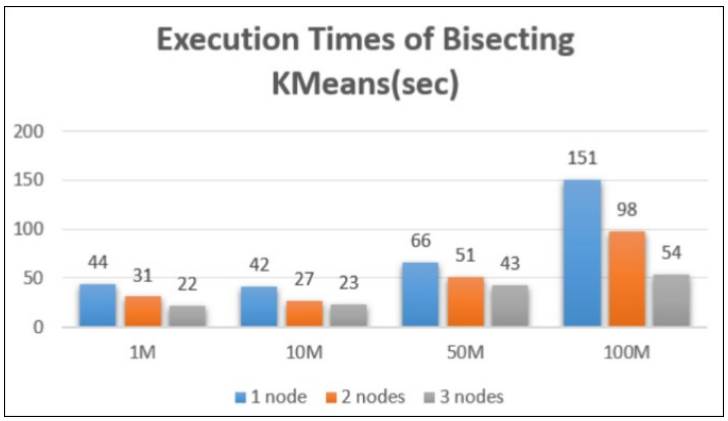


Figure 4.25 pySpark Bisecting-KMeans clustering scalability results

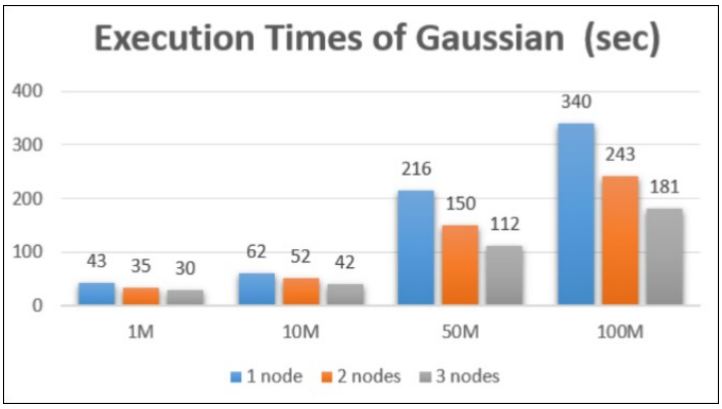


Figure 4.26 pySpark Gaussian Mixture Model clustering scalability results

Based on scalability results using four experiments including three clustering algorithms mentioned above and in addition K-means-RDD mode in Apache Spark clusters. It seems that to add a worker reduces the related clustering's execution time by seconds. As shown in Figures 4.23, 4.24. 4.25 and Figure 4.26, our scalability outputs demonstrate that the bisecting K-means algorithm is significantly quicker than the standard and K-means RDD approaches. When compared to low size data, the effects of scalability are substantially more apparent for higher size data. Due to its process head, the Gaussian mixture model requires more time than the other clustering models. Compared to the standard K-means implementation, the RDD-based K-means implementation produces more effective clustering.

4.2.4.2 Experimental Study for Provenance Data Comparison

This section examines the effectiveness of proposed graph comparison methodology for the DCA-Prov modes, DCAlabel, and DCA_{props}. Figure 4.27 provides an illustration of the example visualization. Table 4.6 shows the accuracy results of the original DCA, Chen's DCA, and the enhanced DCA (E-DCA) DCA_{props} mode that was implemented in this study using test graphs. It is seen that in our DCA_{props} algorithm, compared to the DCA and Chen's DCA, the accuracy is noticeably improved, and we can obtain superior results with our DCA-Prov solution. Performance is affected by the contributions that we provided as discussed in Table 4.2.

Table 4.6 Graph comparison accuracy results

Dataset	DCA Algorithm	Accuracy
10GB noisy data	DCA-Original	88.1
10GB noisy data	DCA-Chen	76.3
10GB noisy data	DCA_{props}	97.3
AMSR-E data	DCA_{props}	96.07

The performance results for the comparison of 10KB, 20KB, and 50KB graphs are shown in Table 4.7. Our performance is reduced to significant levels in our DCA_{label}

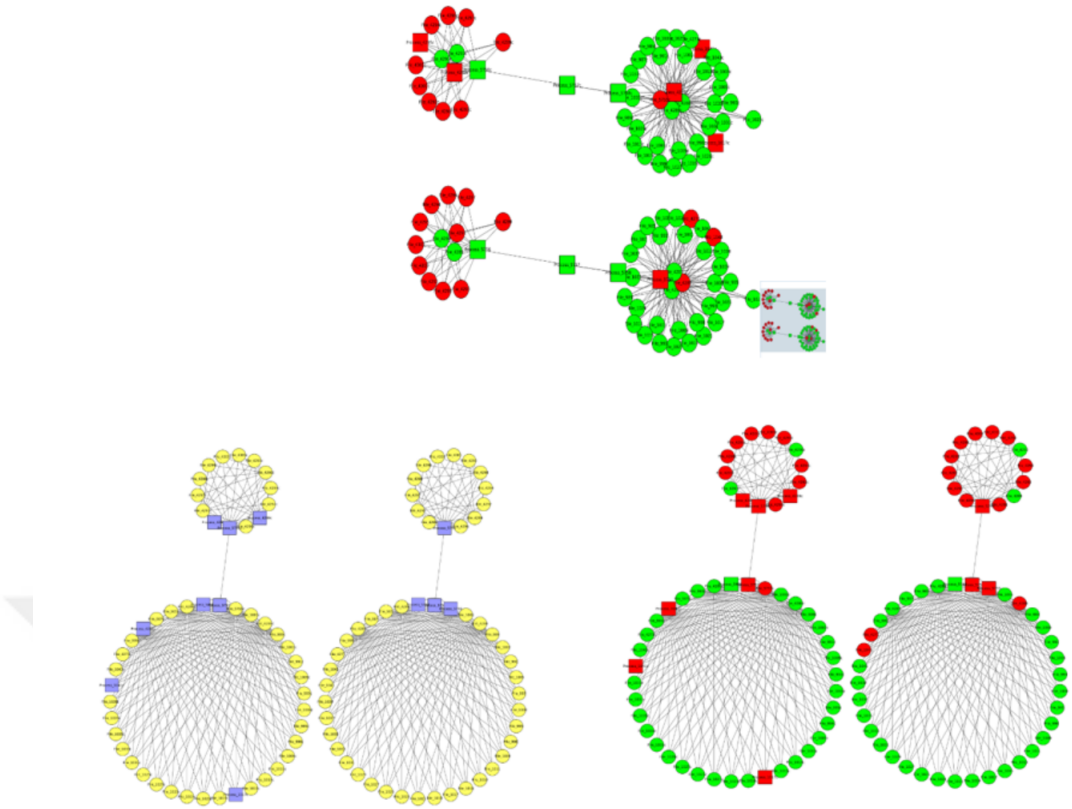


Figure 4.27 DCA_{label} (up) and DCA_{props} before/after graph comparison (down) on sample NASA AMSR-E dataset

Table 4.7 Graph comparison results

DCA Algorithm	Data size(KB)	Performance(ms)
DCA-Original	10/20/50	71/167/7267
DCA-Chen	10/20/50	20/213/5947
DCA_{label}	10/20/50	6/10/277
DCA_{props}	10/20/50	27/92/1861

technique, as can be seen. Furthermore, we achieve better results using DCA_{props} technique than the DCA and Chen's DCA by the given contributions. We increase performance by determining the node type prior to calculating attendance and introduce a minimum threshold value to reduce unnecessary expenses associated with attendance list operations.

4.2.4.3 Experimental Study for Provenance Data Dependency Querying

We used the datasets listed in Table 4.8 for comparison the performance of Chen's BDM (Ctextunderscore BDM) and our upgraded BDM (Etextunderscore BDM). According to BDM comparison data, Chen's BDM technique has improved performance by at most 75 percent, as seen in the comparisons in Figures 4.28, 4.29, and 4.30. We achieve a significant reduction in BDM initialization time, particularly by enabling an improved dependence data structure and data querying technique. The test data used to compare performance consists of various node/edge counts in various sizes, as shown in Table 4.8.The further evaluations on the proposed visualization methodologies's usability is given in Chapter 5.

Table 4.8 Performance data edge and node information

Data size	Node count	Edge count
100 kB	283	552
400 kB	875	2265
1000 kB	1984	5541

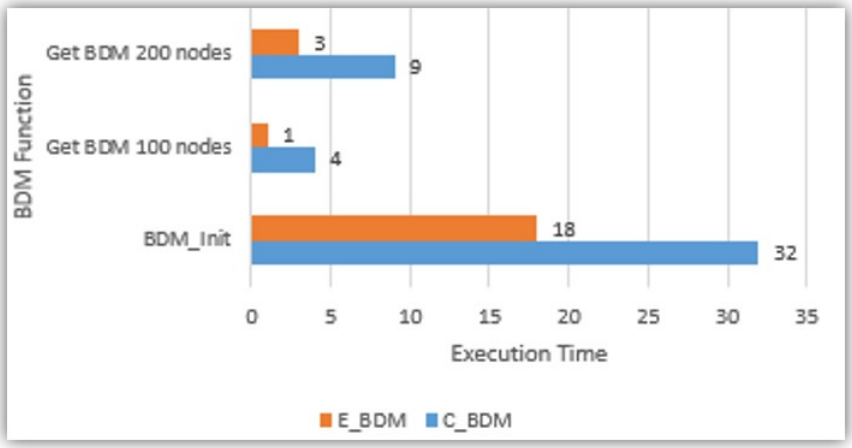


Figure 4.28 100k size E_BDM and C_BDM comparison



Figure 4.29 400k size E_BDM and C_BDM comparison

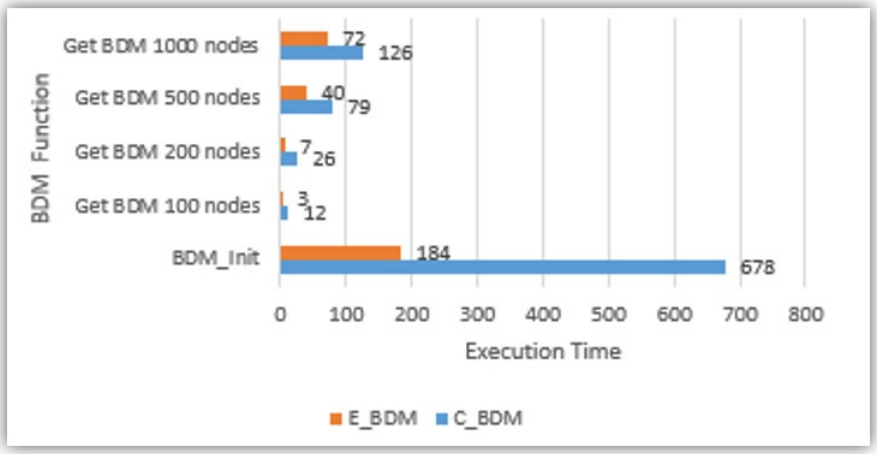


Figure 4.30 1M size E_BDM and C_BDM comparison

4.2.5 Discussion on Research Questions

In this section, we provide discussions on the research questions explained in 4.1.1. in following responding to the questions with the proposed and implemented solutions to meet the visualization requirements.

How to achieve interoperable provenance visualization?: In order to answer this research question, we propose a methodology that was explained in subsection 4.2.1.1 in detail. The ontology selection, the criteria that affected the selection, and the implementation of the ontology model in the system are described.

How to support user-centric visualization framework?: In order to respond to this research question, we outlined a methodology including the usability principles that were followed and the design criteria in subsection 4.2.1.2 in detail.

How to give detailed viewing of node properties?: In order to address this research question, we explained our proposed methodology including the node property detail user interface in subsection 4.2.1.2 in detail.

How to handle provenance graph summarization?: In order to respond to this research question, we propose a comprehensive methodology to generate brief explanations from the large-scale provenance data using the clustering approaches' details in subsection 4.2.1.3 in detail.

How to examine scalability in summarization visual approach?: In order to respond to this research question, we provided an experimental study based on scalability experiments where we distribute the provenance data into many numbers of data nodes then we conducted clustering algorithms by using Apache Spark big data processing platform. The details are given in subsection 4.2.1.3.

How to manage comparison of the experiments/workflows?: In order to answer this research question, we propose a visualization methodology to compare two provenance data by introducing two different modes of comparison in subsection 4.2.1.4.

How to support backward, forward dependency querying?: In order to respond to this research question, we outlined a methodology for dependency querying in backward and forward directions in subsection 4.2.1.5 in detail.

How to provide streaming provenance data visualization?: In order to address this research question, we outlined a methodology to visualize provenance data in a streaming mode in a defined visualization frame in subsection 4.2.1.5 in detail.

4.2.6 Threats to The Validity of the Research Results

We designed and implemented a set of provenance data visualization methodologies that we described in our motivations listed in Section 4.1. In general, we may consider some validity issues and limitations in interoperability, HCI design, summarization, and comparison. The W3C ontology standard, known as the PROV-O model, was chosen as the study's ontology language [52]. However, studies on provenance visualization published in the literature have employed a variety of interoperability groups and solutions.

We placed a lot of emphasis on HCI design and usability since we wanted to conduct the study with the end user at the center. For our HCI design, we used ten well-known Nielsen design principles [63], but there are other concepts and recommendations

found in the literature for usability research. In this study, we proposed a set of conventional unsupervised machine learning algorithms however there are some other deep learning-based clustering and statistical summarizing approaches in the literature to be used in data summarizations solutions.

Although other algorithms are also available for graphical comparison, Chen's technique [40] served as our starting point and primary reference point for the proposed comparison methodology. We then designed and implemented our own solution to make a scientific contribution by enhancing the DCA algorithm.

4.3 Related Work and Our Contributions

The majority of provenance visualization research focuses on graphical displays, graphic comparison methods, and small-size graphics. In order to evaluate provenance traces in a general fashion, PROV-O-Viz presents web application based visualization tool for provenance data provided by various resources [11]. **This study only provides some basic visualization features, however, our study differs from this study by proposing significant visualization methodologies in a comprehensive visualization framework instead of the web application. In addition, user-centric design is one of the goal of our study.**

In an interactive visualization project, the comic strips were produced using the recorded provenance graphs based on the quantified workflows [13].**This study only focuses on a specific usage in health domain however our visualization approaches are more generic ones. Furthermore, our tool has comparison, streaming, and querying approaches in addition.**

Researchers from five different fields can uncover and explain map flaws with the use of the visualization tool Probe-It [14]. **This study differs from these previous works in that it mainly focuses on data visualization techniques such as summarization and comparison of provenance graphs. Furthermore, our proposed visualization approach is not tightly coupled to a workflow management system.**

The prerequisites for the process, timeline, participation, outcomes, interpretation, relationship, and comparison are outlined, and the elements of provenance visualization are identified with reference to Kunde's significant work [15]. Chen [40] satisfies Kunde's list of requirements with customized layout designs and a number of visualization techniques that make it easier to visualize networks in the OPM model. By offering OPM provenance visualization in ordet to conduct network data analysis, Chen collected e-Science related provenance data and put several

data visualization techniques into practice. Additionally, Chen offers a graphical comparison using a sophisticated direct classification technique (DCA) that compares two OPM provenance data graphs for similarity. **This study differs from the study by proposing forward dependency querying, enhancing Chen's backward dependency querying and providing streaming visualization. The DCA algorithm is expanded for comparing two separate provenance graphs in a manner similar to Chen's work but providing two comparison modes. In addition, instead of using OPM we facilitate the PROV-O model specification expression capabilities. Provenance data abstraction is suggested for the limitations and sensitive contents as a part of user tool. For graph summarization,we utilize clustering methodologies for PROV-O data graphs. Additionally, the Cytoscape visualization framework now includes a PROV-O provenance graphic visualization plug-in.**

A study[23] focuses on streaming data applications of diverse active end-users and establishing temporal data streams in IoT domain applications on the reusability for common data streams in the smart city infrastructure rather than reuse processes and outputs after execution. **In our research, similarly we provide streaming visualization using for the PROV Ontology model (PROV-O).In addition our study provide more visualization methodologies such as comparison and history retrieval in a complete visualization prototype.**

MGExplorer[42] provides visual representation of provenance information, combining various visualization approaches and visual querying, while representing provenance information as views. This allows users to retrace their analytical steps and find alternate exploration paths. **Our study differs from this study by providing interoperability using a comprehensive PROV-O Ontology model and providing streaming visualization. In addition we provide a detailed usability study to assess the system's usability.**

KnowledgePearls [71], proposes a method for quickly retrieving analysis states with provenance graph structures that include automatically recorded user interactions and representations. **Our study focuses on user interactions as well but differently, we apply HCI principles for designing and implementing the visualization approaches and the usability evaluations are conducted afterward to assess the usability.**

UVisP [66] was developed to study provenance data, enables users to manipulate and analyze provenance (logs) with implicit prior knowledge of the "Gestalt" HCI theory of perception. **The study focuses on user-centric visualization however, our study differs from this study by proposing summarization, comparison, stream visual-**

ization and detailed investigation methodologies.

Trrack [67] is made to accommodate a variety of use cases, from straightforward action recovery to capturing intent and reasoning, and is designed for simple integration into current or future visualization systems. It can also be used to exchange states with collaborators and keep provenance on a server. **The study proposes summarization methodologies while we propose comparison and streaming visualization methodologies. Furthermore, we support basic visualization features such as sorting, and filtering.**

Prov-Viewer [73] provides a provenance visualization tool that allows users to interact with provenance data exploration using standard features like zooming, filtering, and coloring. **Both studies support graph viewing however our study differs from this study by providing comparison, streaming visualization, and sorting methodologies.**

InProv [72] provides a visualization tool that uses an interactive radial-based tree layout to show filesystem provenance. The tool also makes data exploration more user-friendly by using a new time-based hierarchical node grouping mechanism for filesystem provenance data. **The study assesses the success of the proposed tool by generating a set of question answered with a Likert Scale, however or study is more comprehensive using general SUS metrics besides functional and general usability questionnaires.**

Contributions: This study addresses the Kunde's provenance visualization requirements explained in Section 4.3 as follows. The ontology-based provenance modeling we used for this study addresses the **Process, Results,** and **Relationship** requirements; W3C PROV-O denotes provenance items, activity, entity, and agent as nodes, and their relationship as edges. All intermediate and final outcomes of processes are incrementally included when the PROV-O items are displayed in our visualization environment as a single comprehensive graph. And **Participation** requirement is met by having Agent entity in PROV-O ontology model indicating participants evaluate the correctness of the workflow.

The PROV-O model incorporates node information for nodes and edges as an attribute, such as StartDate and EndDate, to satisfy the **Timeline** requirement. Users can gradually improve their comprehension of the PROV-O ontology model with the Replay and Real Time functionality we propose. For the **Compare** requirement of Kunde, scientific users would like to compare edges' and nodes' properties to compare two workflow provenance graphs. **Interpretation**, in order to support user for the

basic and newly studied visualization methodologies are presented in our study including use cases.

This study proposes a number of provenance visualization approaches so that users can better access, assess, and comprehend provenance data. With the proposed visualization methodologies we aim to design and implement a user-centric, interactive and interoperable data visualization environment. We facilitate the W3C PROV-O ontology specification for provenance data respresentation in oder to support interoperability. Nielsen's ten heuristics are applied to the visualization methodologies to design user-centric HCI elements properly.

The needs for provenance visualization are the subject of study, particularly in the e-Science field. Research on data provenance is crucial for ensuring the reliability and trustworthiness of findings in provenance data produced by workflows. This project focuses on the extraction of provenance data from scientific workflows in graphs and the execution of necessary visualization steps using a complete visualization tool. We offer the summarization strategy in order to provide concise summaries with adequate information so that the enormous bulk of provenance data may be better understood. Additionally, on-demand access to detailed node information is provided by user-friendly HCI features.

To address the demands on the streaming data visualization of scientific workflow operations, we offer a prototype of a streaming visualization of provenance data. By focusing on a provenance entity, the link between PROV-O entities are examined with the use of backward and forward dependency queries to find post-prior relationships. The analysis of sequential scientific operations for anomaly detections requires comparison between provenance graphs, which is a difficult component of knowledge discovery and analytics. We suggest a graphical comparison method for provenance graphs based on PROV-O specifications to identify their similarities and differences with two comparison modes, DCA_{label} and DCA_{props}.In our study, we make use of three dataset: the synthetic social media data, the AMSR-E provenance archive collection from NASA, and a 10 GB semi-synthetic provenance database created at Indiana University. The experimental data shown in the paper demonstrates that the suggested visualization methodologies are workable and satisfy visualization requirements with a negligible overhead, particularly in the e-Science field.

We use a prototype visualization tool to provide the usability of the suggested visualization methodologies. Users of the tool can navigate interactively using common visualization elements. In order to manage the streaming provenance data in PROV-O model, we also design a provenance stream handling and data visualization

technique. With a thorough experimental investigation, we assess the effectiveness and applicability of the suggested visualization methodologies. We also conduct a thorough usability research to assess the prototype's usability from that angle. The findings show that, as explained in Chapter 5, the introduced visualization methodologies have a low processing overhead and they are scalable, offering a high level of user satisfaction. In addition, our study conducted the first in-depth SLR investigation on the visualization of data provenance, with thorough study results given in Chapter 3.

5
USABILITY STUDY ON PROVENANCE DATA VISUALIZATION TOOLS

Several visualization methodologies are proposed during this thesis study as mentioned in Section 4.2.1 to meet the motivations identified in Section 1.1. Since our main goal is to design and implement a set of user-centric and interactive provenance visualization, one of the research questions was, whether the proposed visualization methodologies meet the visualization requirements.

In this chapter, we introduce the conducted thorough usability study on the data provenance visualization domain and present the results in detail at both functional and overall levels. In addition, in order to evaluate our system's usability using a global view and well-known standard, we applied System Usability System (SUS). The detailed usability study experimental study results show the introduced visualization methodologies are usable, and applicable and processing overhead is quite low in the utilization of visualization. We present the Overall Approach, usability Study, and Related Work and Contributions in the following subsections.

5.1 Overall Approach

Implemented visualization prototype including all of the studied visualization approaches is evaluated using well-defined user studies. The participants assessed functional level satisfaction and overall satisfaction by conducting well-defined user studies. In order to take advantage of an industry standard, the participants were also asked to assess the system's usability using SUS metrics.

The user studies per methodology, Provenance Data Dependency Querying, Provenance Data Comparison, Summarization of Provenance Data, and Streaming Visualization are defined in detail explaining the methodology, corresponding user interface, and user tasks. The participants are then asked to finish the tasks and assess the usability of each by answering the questionnaires. Then the results are obtained

from the assessments to find the functional level, general level satisfaction, and SUS scale of the proposed methodologies' usability. The details are given in the following subsections.

5.1.1 Research Questions

We aim to address the following research questions to shape our research for designing and conducting a usability study to meet the usability assessment need of a data provenance visualization tool.

Table 5.1 Research Questions

ID	Visualization Requirement
RQ1	How to design and conduct a usability study on provenance data visualization?
RQ2	How to prepare the user surveys on the provenance data methodologies?
RQ3	How to form the participants?
RQ4	How to establish a usability study environment?
RQ5	How to assess user survey responses?
RQ6	How to assess further comments of the users?
RQ7	How to establish a SUS evaluation for the overall methodologies?

5.2 Usability Study

The details of the conducted comprehensive usability study are given in this subsection. First, we will talk about our methodology, secondly, we will explain the test sessions and the participants. Then we will present the evaluation tasks. Experimental results and discussions will be given and finally, we will talk about the threats and validity of the usability study.

5.2.1 Methodology

Participants in the usability study performed user studies remotely after logging into the testing environment and performing activities as directed on our research website. The test's objective was to evaluate the visualization goals, interface design of the application, flow of information , and information design of visualization approaches for usability.

30 test participants took part in the usability study in the sessions and the length of each individual session was roughly 45 minutes. The introduced visualization techniques were generally agreed upon by all participants to be clear, simple, effective, and easy to utilize. At the end of each user study, the users' were asked for further individual comments and the responses are given in 5.2.6.2.

Each functional group's user tasks are included in the usability study to make it easier to conduct functional and overall satisfaction surveys to assess the suggested visualization prototype. A System Usability Scale study was also established to assess the system using specific SUS metrics.

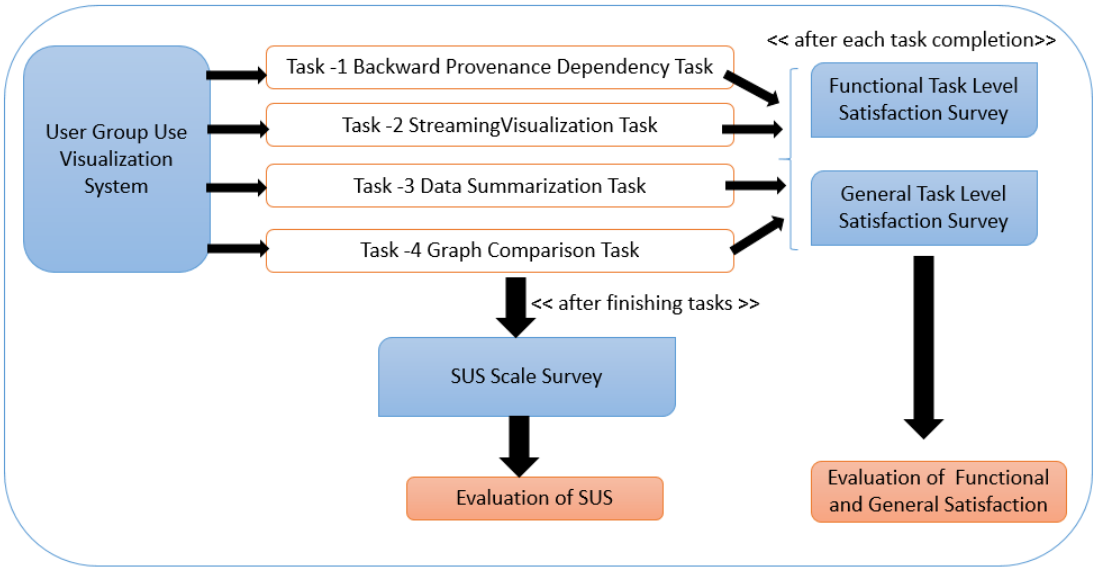


Figure 5.1 Usability study methodology

As illustrated in Figure 5.1, Functional Task Level ad General Task Level satisfaction surveys were conducted after each task definition. In addition after the completion of all tasks, participants complete out the SUS scale survey. Then collecting all the results the corresponding evaluations are done for all Functional Task Levels, General Task Level satisfactions, and SUS evaluation.

This study contains the participant's feedback, satisfaction ratings, task completion and ease/difficulty ratings, errors, and recommendations for improvements. The website for the data provenance usability test is available in the link [87].

5.2.2 Usability Testing Sessions

The test administrator for the usability research recruited participants from potential users among the teammates and coworkers. In order to inform participants of test procedures and to determine their availability, the test administrator sent emails. The length of each session was roughly an hour. The participants were instructed to complete the tasks in accordance with the guidelines provided on the user study website and to complete each visualization task separately during the test session. The test administrator went over the specifics of the test session for each section of the visualization task. The task steps in each segment ended with a request for the

participants to complete a short questionnaire.

The test administrator asked the test participant to score the visualization strategies and user interface design using a five point Likert Scale with measures ranging from Strongly Disagree to Strongly Agree after each task compilation. Two separate usability metric groups are the subject of evaluation in the task's questionnaires for each visualization user study group on the investigated data visualization methodology. Functional Task Level Satisfaction for visual methodology-specific usability evaluation precisely defined for each work group. Overall Task Level Satisfaction for general usability assessment including the same question set for each task group including Ease in Finding Information, Design Simplicity, Task Completion Effectiveness, Flexibility, System State Monitoring, and Overall System Satisfaction.

5.2.3 Usability Test Participants

All the test participants were possible users for the provenance data visualization system. Most of the participants were from Information Technology (IT) community members from a research and development institution. All the participants are involved in testing between December 21th – 27th, 2021. 30 participants, 13 were male and 17 were female and the distribution is shown in Table 5.2.

Table 5.2 Usability study test participants

IT/Engineering	**R**esearch/Institution	**O**ther
19	6	5

5.2.4 User Surveys and Evaluation

Test participants were asked to complete the following user studies per methodology and conduct the user surveys afterward a- finding backward data dependency for a specific provenance visualization item,b- observing streaming visualization of provenance data items and monitoring the entities' arrival times, c- creating data summarizations of provenance data, d- comparing two provenances in two comparison modes.

The Likert scale had a maximum score of 5(Strongly agree) and a minimum score of 1(Strongly disagree). Theorem 5.1 is used to get the mean agreement value. A mean agreement rating of > 4.0 is taken into account as the user agreeing that the information was simple to find, that they could keep track of their location, and that they could successfully complete the tasks that were assigned. Agree ratings are the agree and strongly agree on ratings combined.

Theorem 5.1. *Mean Agreement Calculation*

for each scale in Likert Scale:
total weight *+= number of people selected response * weight of response*
end

mean agreement *=* ***total weight*** */ total number of participant*

5.2.4.1 Functional Level Task Definitions

After the completion of each task on a specific visualization methodology, test participants rated their task satisfaction with respect to the functional level aspects of each provenance visualization approach as listed in the following. The complete set of user survey questions is given in Appendix-A User Scenarios and Tasks.

Task-1: Backward Provenance Dependency Visualization Methodology: Backward provenance data dependency methodology based usability aspects are identified to see if the end-user: a- can quickly focus on any visualization component to efficiently query its backward dependency, b- can clearly trace the relationship between input and output nodes to examine the dependency between them.

Task-2: Streaming Data Visualization Methodology: Streaming visualization related usability aspecty are defined to assess the methodology's efficiency consisting: a- the capacity to clearly track the arrival of data items at their creation time, b- data sequence analysis and online event tracking.

Task-3: Provenance Data Summarization Methodology: Summarization methodology-related functional level usability factors are given to evaluate summarization requirements to see whether: a- users are assisted by the summarization using data abstraction to give simplicity, b- instead of using enormous amounts of data, online event tracking focuses on a specific data group.

Task-4: Provenance Data Comparison Methodology: Provenance data comparison analysis' usability factors are identified below to assess proposed comparison modes to check if: a- the needs for comparison are satisfied, b- the application assists in identifying data components that are similar and those that are not, c- the effective configuration of comparison parameters through the user interface, d- by performing further comparisons, the display facilitates comparisons between various setup options.

5.2.4.2 General Level Task Definitions

General task level satisfaction metrics, Ease in Finding Information, Design Simplicity, Effective Task Completion, Flexibility, Feedback, and System Satisfaction related questions were asked to the participants at each task completion in order to assess general user interface HCI features. The complete set of user survey questions is given in Appendix-A Usability Study.

5.2.5 Experimental Study Results

In the following subsections, we present the Functional Task Level satisfaction in graphics for each visualization methodology and afterward the General Task Level satisfaction results.

5.2.5.1 Functional Task Level Satisfaction

After the execution of the user studies for the studied methodologies, the defined user surveys took place and the users were asked to give their responses on their satisfaction on a functional level. The survey results will be given in the following.

Task-1: Backward Dependency Querying Visualization: Functional task level satisfaction on Task-1 evaluation results are given in Table 5.3 according to functional aspects and overall value. The results show that the overall Agree rating meets the limit condition (>4) as described in Section 5.2.4 which means the user agrees on the usability and applicability. In addition, the scatter diagram in Figure 5.2 illustrates the responses of 30 participants individually, for each functional aspect as well.

Table 5.3 Functional Level Usability Task 1 – Mean Task and Percent Agree Ratings

Functional/ Task	Focus to Any Data Item	Relational Analysis	Post-Prior Analysis	Overall
Implement / Backward Dependency Querying	4.53 (100%)	4.43 (100%)	3.96 (70%)	4.3

Percent Agree % = Agree and Strongly Agree responses combined.

Task-2: Streaming Data Visualization Methodology: Functional level task satisfaction on Task-2 evaluation results are given in Table 5.4 according to functional aspects and overall value. The results show that the Agree ratings for all aspects meet the limit condition (>4) as described in Section 5.2.4 which means the user agrees on the usability and applicability. In addition, the scatter diagram in Figure 5.3 illustrates the responses of 30 participants individually, for each functional aspect.

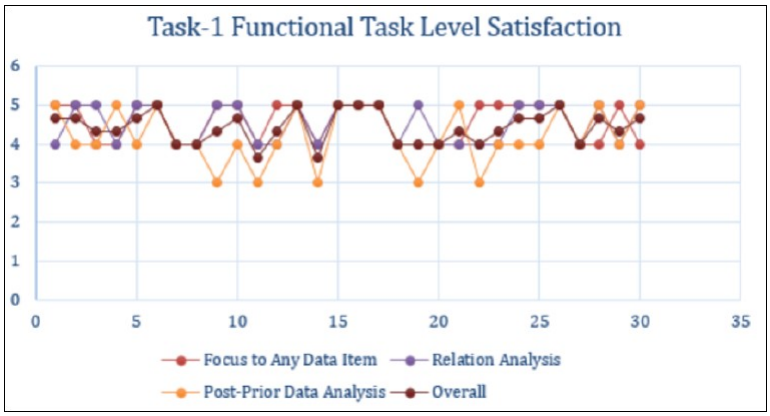


Figure 5.2 Task-1 Usability Test Responses

Table 5.4 Functional Level Usability Task 2 – Mean Task and Percent Agree Ratings

Functional/ Task	Data Exploration	Sequence Analysis	Online Event Tracking	Overall
Implement Streaming Visualization steps	4.56 (100%)	4.53 (100%)	4.2 (83%)	4.43

Percent Agree % = Agree and Strongly Agree responses combined.

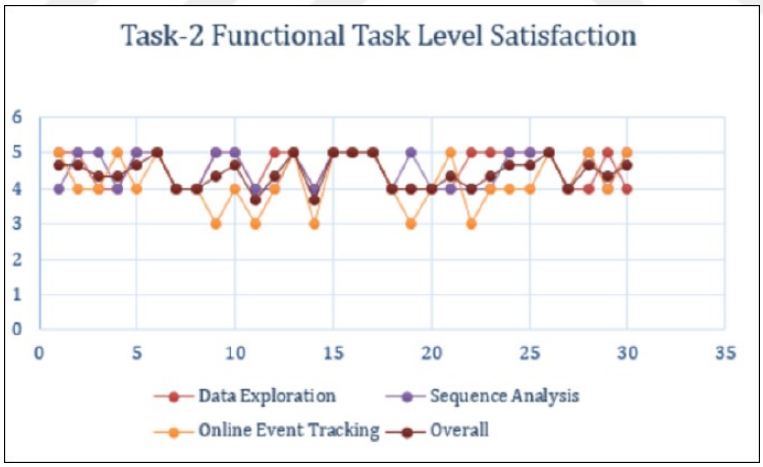


Figure 5.3 Task-2 Usability Test Responses

Task-3: Provenance Data Summarization Methodology: Functional level task satisfaction on Task-3 evaluation results are given in Table 5.5 according to functional aspects and overall value. The results show that the Agree ratings for all aspects meet the limit condition (>4) as described in Section 5.2.4 which means the user agrees on the usability and applicability. In addition, the scatter diagram in Figure 5.4 illustrates the responses of 30 participants individually, for each functional aspect as well.

Table 5.5 Functional Level Usability Task 3 – Mean Task and Percent Agree Ratings

Functional/ Task	Simplicity	Focus to the Relevant Data	Data Separation	Overall
Implement Data Summarization steps	4.33 (93%)	4.53 (100%)	4.46 (96%)	4.44

Percent Agree % = Agree and Strongly Agree responses combined.

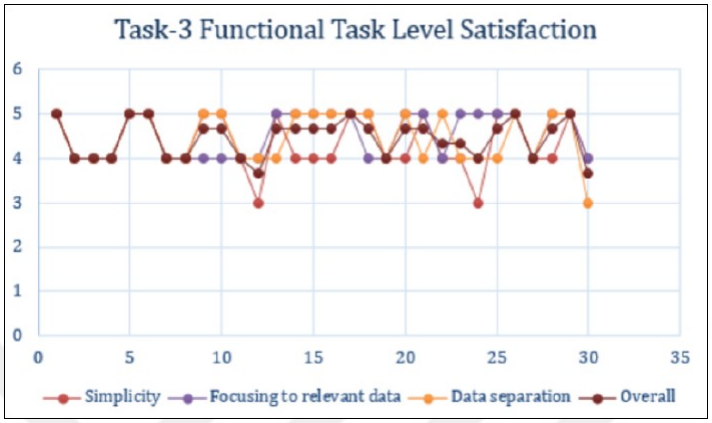


Figure 5.4 Task-3 Usability Test Responses

Task-4: Provenance Data Comparison Methodology: Functional level task satisfaction on Task-4 evaluation results are given in Table 5.6 according to functional aspects and overall value. The results show that the Agree ratings for all aspects meet the limit condition (>4) as described in Section 5.2.4 which means the user agrees on the usability and applicability. In addition, the scatter diagram in Figure 5.5 illustrates the responses of 30 participants individually, for each functional aspect as well.

Table 5.6 Functional Level Usability Task 4 – Mean Task and Percent Agree Ratings

Functional/ Task	Simplicity	Distinctiveness	Customizability	Successive Comparison	Overall
Implement Data Comparison steps	4.53 (100%)	4.63 (100%)	4.46 (100%)	4.3 (93%)	4.48

Percent Agree % = Agree and Strongly Agree responses combined.

5.2.5.2 General Task Level Satisfaction for Visualization Methodologies

General task level satisfaction metrics, Ease in Finding Information, Design Simplicity, Effective Task Completion, Flexibility, Feedback, and System Satisfaction related questions were asked to the participants at each task completion in order to assess general user interface HCI features. Table 5.7 depicts the Mean Task Ratings and Percent Agree on the results of each aspect and the overall value for each task. The

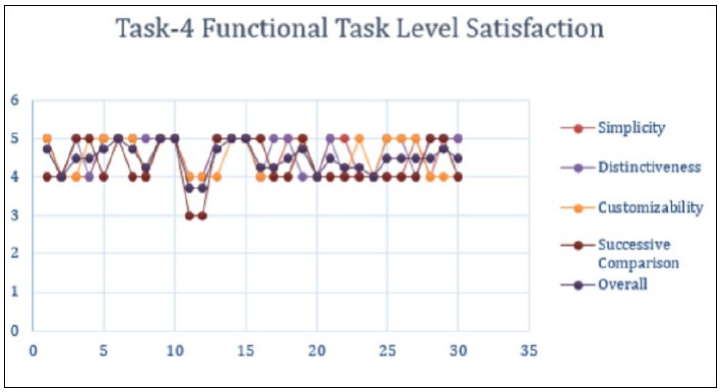


Figure 5.5 Task-4 Usability Test Responses

results show that the Agree ratings for most of the aspects and all overall values meet the limit condition (>4) as described in Section 5.2.4 which means the user agrees on the usability and applicability. In Figure 5.6, 3D bar chart illustrates the responses for the general task satisfaction survey per task. In addition, the scatter diagrams in Figure 5.7 illustrates the responses of 30 participants individually, for each general aspect.

Table 5.7 General Usability – Mean Task and Percent Agree Ratings

Task	Ease in Finding Information	Design Simplicity	Effective Task Completion	Flexibility	Feedback	System Satisfaction	Overall
Task-1 Implement Bacward Dependency steps	4.46 (96%)	4.13 (90%)	4.33 (96%)	4.1 (80%)	3.8 (60%)	4.32 (96%)	4.2
Task-2 Implement Streaming Visualization steps	4.4 (96%)	4.2 (88%)	4.3 (90%)	4.3 (90%)	3.85 (65%)	4.3 (95%)	4.2
Task-3 Implement Data Summarization steps	4.5 (96%)	4.46 (95%)	4.26 (96%)	4.1 (86%)	3.9 (63%)	4.4 (98%)	4.27
Task-4 Implement Data Comparison steps	4.5 (96%)	4.3 (96%)	4.3 (93%)	4.1 (83%)	3.96 (66%)	4.45 (100%)	4.26

Percent Agree % = Agree and Strongly Agree responses combined.

5.2.5.3 SUS Evaluation of Overall Methodologies

The usability testing participants were asked to assess overall methodologies according to the SUS metrics which are explained in Subsection 2.4.1. The SUS metric responses of the participants are given in Table 5.8. The radar diagram in Figure 5.8 display the SU scale mean rating and percent agree on values of the responses.

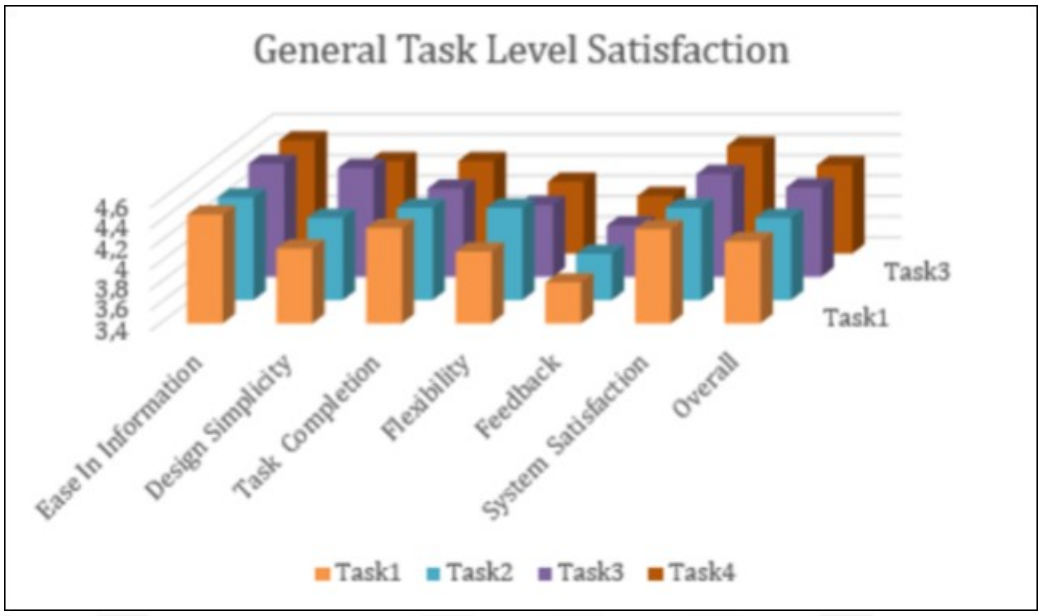


Figure 5.6 General Task Level Satisfaction Diagram

Table 5.8 SUS Evaluation – Mean Task and Percent Agree Ratings

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Mean Rating	Percent Agree
1- I think that I would like to use this system frequently.	-	3	6	14	7	3.83	(70%)
2- I found the system unnecessarily complex.	6	19	3	2	-	2.03	(7%)
3- I thought the system was easy to use.	-	-	4	19	7	4.1	(86%)
4- I think that I would need the support of a technical person to be able to use this system.	3	17	6	4	-	2.36	(13%)
5- I found the various functions in this system were well integrated.	-	-	3	16	11	4.26	(90%)
6- I thought there was too much inconsistency in this system.	10	15	3	2	-	1.9	(7%)
7- I would imagine that most people would learn to use this system very quickly.	-	1	4	18	7	4.03	(83%)
8- I found the system very cumbersome to use.	6	17	3	4	-	2.16	(13%)
9- I felt very confident using the system.	-	1	5	20	4	3.9	80%
10- I needed to learn a lot of things before I could get going with this system.	8	13	3	6	-	2.23	(20%)

Percent Agree % = Agree and Strongly Agree responses combined.

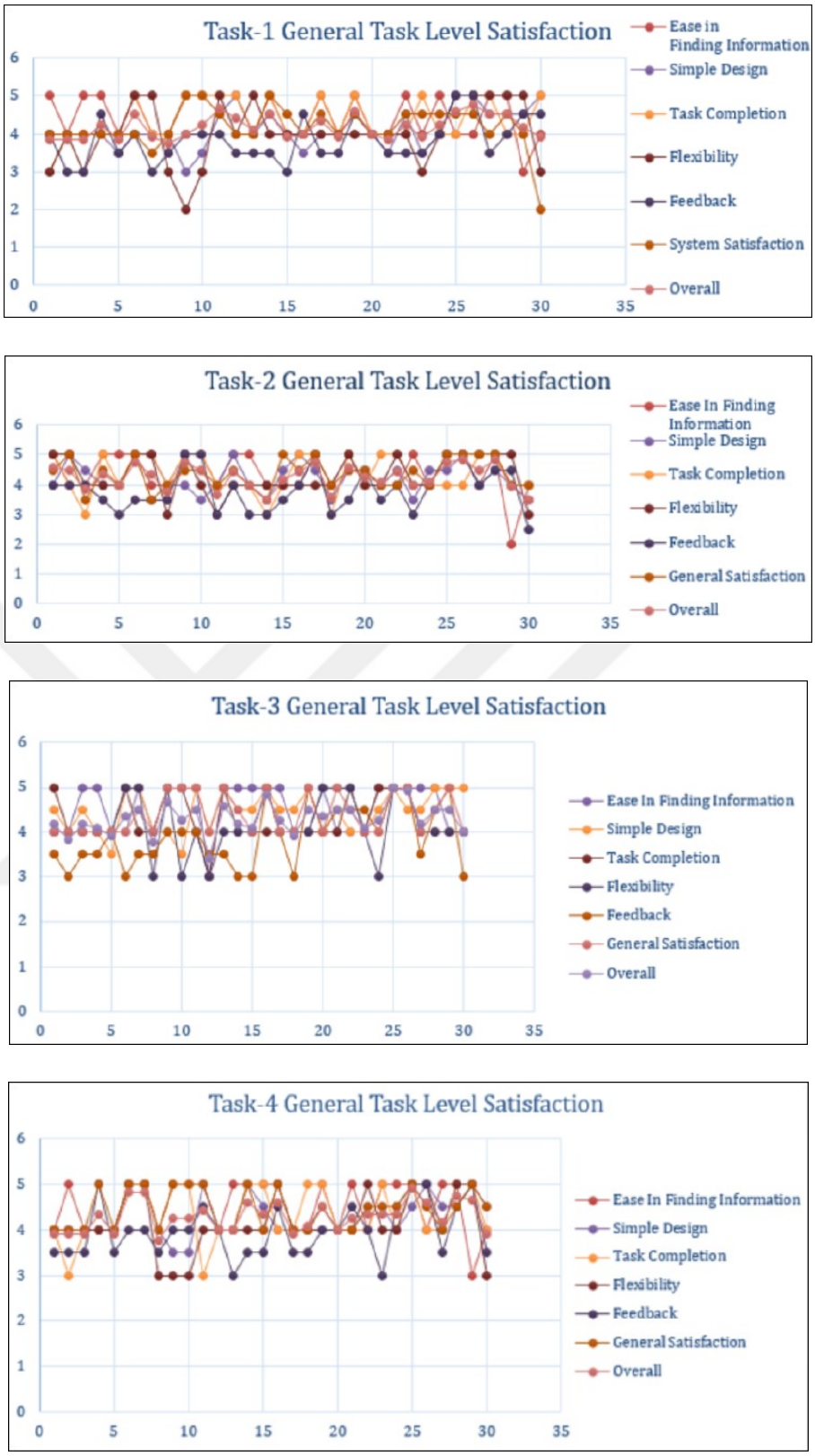


Figure 5.7 General Task Level User Satisfaction for Each Task



Figure 5.8 SUS Mean Rating(left) and Percent Agree(right)

5.2.5.4 Time on Task :

The goal of the usability study was to evaluate our application's four different visualization strategies and user-centered interaction design. After completing each activity, the time spent on it is estimated using the manually recorded submission time. Participants start time is also manually recorded.

The average time spent on each assignment shows that some tasks were inherently harder to execute than others, according to the task's visualization approach characteristic and specific steps. For example, Comparison Task (task-4) has 2 different modes, and participants were asked to try out both modes and configurations therefore it takes longer compared to others. The minimum, maximum, and average task time is given in the Table 5.9.

Table 5.9 Tasks and the durations

	Min-Time	Max-Time	Average-Time
Task1	06:43	12:34	08:35
Task2	08:06	15:37	11:26
Task3	08:58	19:41	13:23
Task3	09:09	22:42	16:29

5.2.6 Discussion on Research Questions

In this subsection, we provide discussions on the research questions explained in 5.1.1. in following.

How to design and conduct a usability study on provenance data visualization?: In order to answer this research question, we propose a methodology that

was explained in subsection 5.2.1 in detail. We established functional task level and general level surveys to assess user satisfaction. In addition, SUS was evaluated for the visualization system.

How to prepare the user surveys on the provenance data methodologies?: In order to respond to the research question, we outline the user survey aspects in Section 5.2.4.1 for the functional task level and in Section 5.2.4.2 for the general level. The complete questions and surveys are given in Appendix-A User Scenarios and Tasks.

How to establish a usability study environment?: In order to address this question, we propose a definition of sessions identified in Section 5.2.2. The usability environment set-up is defined in the research webpage [87] in detail.

How to form the participants?: In order to answer this research question, we outlined the participation details in Section 5.2.3. Participants were selected from the researcher in the institution or from the IT experts who are the possible users of an e-Science visualization system.

How to assess user survey responses?: In order to respond to this question, we explain the assessment details in the evaluation part given in Section 5.2.4. Functional satisfaction and general satisfaction are evaluated differently than the SUS metric calculations. The functional task level and general level experimental results and outputs are given in Section 5.2.5.

How to establish a SUS evaluation for the overall methodologies?: In order to address this research question, we explained the SUS evaluation details in Section 5.2.5.3. Mean agreement and percentage agree calculations were calculated and illustrated individually in the study.

How to assess further comments of the users?: In order to respond to this question, we propose further user comments in Section 5.2.7. The error prevention details and the test participants' further comments on likes, dislikes and recommendations will be considered for future work.

5.2.6.1 Error Prevention :

In our investigation, we anticipate seeing how well our visualization methodologies work for interactively monitoring and controlling provenance data to satisfy visualization objectives. Error Prevention is used in the program by making use of Nielsen's one of the ten heuristics. Visualization The user interface and functionality of the prototype were created to prevent users from making crucial mistakes, including

providing a negative cluster number for summarization or a weight or threshold value outside of the [0-1] range for the comparison setup.
Some non-critical unexpected situations such as low response time in selecting a visualization node or application UI items for some users, due to remote control on the application were faced but these rare situations didn't prevent the successful completion of the scenario. In addition, the rollback mechanisms are also set in the application for undoing the selected function or reloading the visualization data network from scratch easily in one step instead of loading node, edge, and style data from the beginning.

5.2.6.2 Likes, Dislikes, Participant Recommendations

Upon completion of the tasks, participants provided feedback on what they liked most and least about the application, and recommendations for improving the visualization system.

Liked Most: The following comments capture what the participants liked most: a- visual styles selections like color and shape in the Visualization Frame user interface, b- UI items are user-friendly to understand the functionality, c- the design of UI items was organized well.

Liked Least: The following comments capture what the participants liked the least: a- in Task-2, the activation button could be removed to activate the toolbox, replay toolbox could be activated with the first trigger, b- in Task-3, the user could open more than one cluster at the same time.

Recommendations for Improvement: The test participants' comments are taken and considered for future improvements of the system as follows: a- in the Backward Dependency task, there could be an option to sort the dependent nodes in time in order to increase readability, b- it could be a technical contribution as well but similar to Backward Dependency, there could be another approach for Forward Dependency Querying, c- more popular look and feel styles can be adapted to the system instead of conventional ones, d- there can be some additional UI items like Stop button for streaming Visualization task similar to Pause, e- there could be time differences between nodes in the visualization frame for Task-2 streaming visualization, f-the filtering mechanism could be helpful for Task-2 streaming visualization to be able to selective while displaying items according to properties like node-id or relation-type.

5.2.7 Threats to The Validity of the Research Results

We conducted a usability study to prove our studied visualization methodologies' usability and applicability as one of the motivations that we identified in Section 4.1. In general, we may consider some validity issues and limitations in participants, usability surveys, and usability assessment metrics.

Although we have included competent people who work in the academic field and as potential users of the proposed visualization system in the usability study, there are also different potential users who work with scientific workflows and need visualization directly.

We have created a detailed user survey for asking the end-user functional task level satisfaction and general task level satisfaction for the implemented visualization approaches, but maybe we may not have grasped all the necessary questions to measure functionality.

We benefited from SUS, which is used for measurement in the literature, but there are different metrics used in usability studies to assess usability.

5.3 Related Work and Contributions

Usability is often defined as considering how and why end-users use a product, assessment, usability, and user-centered design; as a result, it is critical to gain users' pleasure when their objectives, mental models, tasks, and requirements are all satisfied.

In their research, Oakley-Girvan et. al[92] develop, implement and evaluate a mobile device-based survey system for behavioral risk assessment. They evaluated its feasibility, usability, acceptability, and validity in a usability study including 27 question-set in 3 surveys with 670 participants. **Our study differs from this study by having a small set of participants but following a rigorous usability study assessing the SUS metric of the visualization system.**

Another study [93] focuses on evaluating the accessibility and usability of online employment application Web sites in eight southeastern US states. Web Content Accessibility Guidelines (WCAG) were followed while designing the significant processing. **However, we applied Nielsen heuristics for designing user interfaces in our desktop application instead of the web application.**

In a study [94], heuristic evaluation was conducted in a usability study to assess

the usability of two selected virtual museums offering virtual tours. The usability assessment was conducted having Nielsen's heuristic analysis and a System Usability Scale (SUS) survey. **In our study, similarly, SUS survey was applied for assessing the system usability and in addition a comprehensive usability study including usability surveys in the data provenance visualization area. Nielsen principles were followed only in HCI design.**

Contributions: We describe the in-depth usability study on data provenance visualization and show the results in detail at both the functional task level and general task level. Additionally, we used the System Usability System to evaluate the usability of our tool from a broad perspective and using a recognized standard (SUS). The thorough usability study experimental results demonstrate that the suggested visualization methodologies are applicable, usable, and have a very low processing overhead that makes the visualization system easier to use.

A rigorous approach was followed for the usability study. Usability study environment set-up instructions, visualization methodology-related usability task definitions, and user surveys are utilized in a usability study environment. Participants' feedback for functional task level, general task level surveys, SUS survey, and completion time information were collected by the environment in order to assess the usability of the system. Furthermore, additional feedback was get from the test participants as a free style form. After the usability tests, all the further feedback (in the heading Likes, Dislikes, Participant Recommendations) of the test participant will be taken into account carefully to enhance the Least Liked ones and to assess recommendations in order to improve overall ease of use and address the areas where participants experienced unexpected behaviors or found the interfaced/information design unclear. The recommendations will be taken into account according to the assigned rank value.

6
RESULTS AND DISCUSSION

Research on data provenance is crucial for ensuring the reliability and trustworthiness of discoveries in several domains including application use domains identified in this study in Section 1.3.2. This study focuses on the extraction of provenance data from application use domains and propose a number of provenance visualization methodologies allowing users to obtain, assess and comprehend provenance data efficiently. Here we present a summary of the study responding the research questions identified in Section 1.2.

How can we design and implement a user-centric,interactive and interoperable provenance visualization tool including visualization approaches?

In order to achieve interoperability for provenance information exchange in visualization, we want to make use of well-defined web semantics, including data models and definitions. Three basic ontology item kinds are necessary for a comprehensive provenance data specification. Things can be either physical, digital, or of other forms. Activities can use existing entities or generate new ones. Agents can take part in activities and be held accountable. Entity objects are things of any kind.

In this study, the provenance objects must be identified using a clear web semantic use in order to facilitate interoperability as a quality feature. The W3C PROV-O ontology model is chosen for this study because it contains a thorough model definition and is being utilized as a model for related research in application use areas. It adheres to the guidelines for data management in the permitted application domains.The PROV-O Ontology satisfies the requirements for provenance representation and offers a high level of interoperability for data interactions because to its extendable structure. Hence the provenance data is imported in the W3C PROV-O ontology format combination of W3C PROV-O entity, agent and activity components in order to support interoperability [39].

In order for end users to better understand provenance data, we aimed to investigate a set of visualization methodologies. In order to achieve this goal, visualization research should take user navigation models and HCI principles into account. To make user navigation as simple as possible, the fundamental components are developed and put into use on an open source visualization framework. The ten Nielsen heuristics are followed by us: (1) visibility of system state, (2) use of metaphors, (3) freedom and control of the user, (4) consistency and standards, (5) error prevention, (6) recognition rather than recall, (7) adaptability and effectiveness of use, (8) aesthetics and minimal design, (9) assist users identify, diagnose, and recover from problems, (10) help, and documentation, and we present user interface solutions that take advantage of each of these heuristic ideas when creating the HCI of the proposed system [39].

We construct a prototype visualization tool to ensure the usability of the suggested visualization approaches. Users of the tool can do interactive navigation using common visualization elements. In order to manage the online streaming provenance data in PROV-O model, we also design a provenance processing and visualization technique for the streaming data. With a thorough experimental investigation, we assess the suggested visualization methodologies. Additionally, we perform a thorough user study to assess the visualization prototype's usability. The findings show that the suggested visualization methods are scalable, have no processing overhead, and offer high levels of user satisfaction. We intend to look into other provenance graph data comparison and summarizing methods in further studies. Additionally, we intend to expand the provenance dataset in size [2].

How can we design and implement a visualization methodology that can group together data provenance to help end-users better understand the data?

Making concise summaries to aid in a better understanding of data attributes is vital when dealing with large-scale provenance data graphs. In order for the user to immediately focus on the specific groupings, it is helpful to organize the enormous amounts of data into a specific set of objects based on their qualities. The task of grouping various data objects into categories of comparable objects is known as clustering. One group refers to a collection of data. Based on similarity, data sets are split into various categories in the cluster analysis[39].

A property-based clustering is employed in this study to create a clear representation of the similarities between node attributes. To build data groups from the large-scale provenance data listed, three different traditional unsupervised clustering algorithms are investigated: centroid-based clustering, hierarchical-based clustering, and distribution-based clustering. For each method, a representative algorithm—the

K-Means algorithm for centroid-based clustering, the Bisecting K-Means algorithm for hierarchical-based clustering, and the Gaussian Mixture Model for distribution-based clustering—was chosen. The methods were used to implement data summarizing methodologies, and thorough scalability research was also carried out to evaluate the clustering outcomes with the larger provenance data. The findings of the experiments and the usability assessment demonstrate that the proposed clustering methodology provides concise summaries with adequate information so that the enormous bulk of provenance data may be better understood. Additionally, on-demand access to the detailed information is provided by user-friendly HMI capabilities [45].

How can we design and implement a visualization methodology that can help end-users identify the differences and similarities between the two provenance data?

One of the primary functions of data provenance analysis is to determine the differences and similarities between two data graphs for the purpose of anomaly discovery or troubleshooting. In this study, we propose two distinct comparison modes: a simple comparison(DCA_{label}) and an advanced comparison(DCA_{props}), to satisfy users' requirements for comparison visualization for the provenance graphs in PROV-O ontology model to find the differences and similarities between them. An enhanced DCA algorithm was implemented by adding features to the approach and by making the comparison process highly configurable by the user. The experimental studies show that the proposed visualization methodologies can be applied and can satisfy visualization requirements with negligible overhead, particularly in the application domains[45].

How can we design and implement a visualization methodology that can help end-user to understand the dependencies in provenance data?

Metadata can provide some unexpected or undesired results when a data sequence is performed, which are then shown as output. The user wants to look into the input or any intermediary artifacts that contributed to that condition. For the provenance analysis, it is necessary to look at the prospective nodes that might be touched by the ensuing unanticipated or unpleasant events. The data provenance visualization systems should include backward and forward dependency querying to understand dependencies. To prove a causal connection between the PROV-O nodes and their connections, additional investigation and analysis of the PROV-O data are necessary.

In this work, we present an improved Backward Dependency Querying (BDM) to give users with problem-solving capabilities and help them understand how a specific outcome developed in a relationship between input data, intermediate data, and

output to pinpoint process failure steps. To make the PROV-O data dependence analysis easier, a querying approach is also introduced for the PROV-O items' forward dependency in addition to the proposed backward provenance dependency technique. Examples of forward provenance dependence include the analysis of the output and intermediate nodes depending on a selected input node and the subsequent history of this particular node. When querying the future using our PROV-O model, the output, intermediate, and input nodes are composed of activity, agent, and entity elements [45].

How can we design and implement a visualization methodology that can help end-user to analyze streaming provenance data?

The traditional methods for provenance data visualization entail offline processing and storage of the data, which uses up a lot of non-essential resources. Large-scale, protracted scientific projects can generate a sizable volume of difficult-to-remember provenance data. In this article, we'll talk about a streaming visualization in which provenance data activity, entity, and agent pieces are shown with a limited amount of memory within the frame of the visualization and in the order that they arrive.

In this study, we propose an online streaming visualization solution to enable distributed processing of streaming provenance data. By utilizing a publish-subscribe messaging system to get streaming notifications from various distributed settings, the provenance data elements are gathered and saved in in-memory storage to cache the visualization data for presentation in our visualization environment[45].

How can we conduct a comprehensive literature review investigating the significant studies on Data Provenance Visualization in literature with a rigorous process?

It was investigated that the literature review on data provenance visualization is insufficient. It was noted that a thorough SLR study was not included within the scope of this study in order to guide the research on this issue, despite the fact that there are some SLR studies on the topic, they were only included as a minor component of some comprehensive studies.

A comprehensive SLR study was done in the scope of the thesis study which can be regarded as the first thorough SLR study in this field. Each research topic offers data provenance visualization options with a variety of general or niche visualization techniques. The SLR study investigates 18 relevant papers between 2012 and 2022. The research has demonstrated that provenance visualization is one of the most important areas of study to address the needs of metadata analysis, particularly in the e-Science domain. Since every study suggests future plans highlighting the various

unanswered questions that still need research in this field, it appears that the area is still being investigated [46].

A comprehensive SLR study was done in the scope of the thesis study which can be regarded as the first thorough SLR study in this field. Since it proposes a large systematic evaluation of the literature, this work is being recognized as a reference for assessing specific studies on data provenance visualization approaches and technologies.The SLR study investigated 18 relevant papers between 2012 and 2022. The common goal of this research to present an SLR to explain works that explicitly address data provenance visualization in general or in other fields, specifying visualization approaches and techniques. By following a stringent approach, a thorough screening of primary research in the literature was made possible. We undertook an SLR study on data provenance visualization in this study to close the knowledge gap by satisfying the demand for a thorough literature assessment in this field.

How can we conduct a comprehensive usability study on provenance visualization tools to understand the usability from the end-users point of view?

The research on the data provenance visualization approaches draw attention to the importance of usability studies on the visualization methodologies.This study aimed to ask whether the proposed visualization methodologies satisfy the visualization criteria because our major objective is to build and construct a collection of user-centric and interactive provenance visualizations.

We introduced the in-depth usability study on data provenance visualization methodologies that we proposed and showed the results in detail at both the functional and general levels. Additionally, we used the System Usability Scale(SUS) to assess the usability of our system from a broad perspective and using a recognized standard. The thorough usability study experiments demonstrate that the suggested visualization methodologies are applicable, usable, and have a very low processing overhead that makes the visualization system easier to use [43].

6.0.1 Future Research Opportunities

The need for user-centric, interactive, and interoperable data provenance visualization approaches to be provided in a visualization framework was thereby growing. Particularly, provenance data areas like E-Science and Social Media are considered to have a rising need for solutions for data summarization, data comparison, dependency querying, and stream visualization. By adhering to the W3C PROV-O provenance

ontology standard, we devised and implemented visualization approaches to close this gap. To demonstrate that the suggested approaches work as predicted, we conducted an extensive usability research to evaluate them. A thorough SLR analysis was also performed to examine the earlier research in this field and provide guidance for next research.

In this study, we investigated provenance data summarizing methods using conventional unsupervised machine learning clustering algorithms. We believe that supervised approaches to data graph summarization, such as Graph Neural Network (GNN) deep learning algorithms and statistical methodologies, will result in a better comprehension of the data grouping. This is thought to be a subject worth exploring further. Although the study took a user-centric and interactive approach, Nielsen heuristics were utilized as a guide. However, there are other topics in this field. In order to improve usability, we think that visualization approaches can also benefit from applying Ben Shneiderman's The Eight Golden Rules in addition to Nielsen Heuristics. In addition, the further comments of the participants of the conducted usability study will be used to improve the visualization methodologies.

Additionally, as we have so far used unsupervised methods for analysis, obtaining insight may prove challenging for future development. Through supervised approaches to processes like neural network models, it is possible to predict early prevention from false instances or interruptions. It is also feasible to ensure that additional domains, such as network systems or the health domain, can be developed in addition to application use domains. This desktop application can also be made into a mobile or web-based version in the future.

To further successfully test the usability of the methodologies created by this study and needed in the application use domains, other usability studies can be planned. For instance, the direct contribution of one group performing the assigned activities manually and the other performing them while using the provided visualization tool might be used to demonstrate the direct contribution of the provenance data visualization methodologies studied in this thesis. In addition, one of our goals is to involve application-specific experts in the usability study and user testing for further improvements.

REFERENCES

[1] "The fact of coming from some particular source or quarter, source, derivation," in *http://en.wikipedia.org/wiki/Provenance*, Oxford English Dictionary (OED), 2017, April 20, pp. 935–942.

[2] I. M. Yazici, E. Karabulut, M. S. Aktas, "A data provenance visualization approach," in *2018 14th International Conference on Semantics, Knowledge and Grids (SKG)*, 2018, pp. 84–91. DOI: 10.1109/SKG.2018.00019.

[3] R. Hasan, R. Sion, M. Winslett, "Preventing history forgery with secure provenance," *ACM Transactions on Storage*, vol. 5, Dec. 2009. DOI: 10.1145/1629080.1629082.

[4] B. Glavic, K. Dittrich, "Data provenance: A categorization of existing approaches," in *Datenbanksysteme in Business, Technologie und Web (BTW 2007) – 12. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS)*, A. Kemper *et al.*, Eds., Bonn: Gesellschaft für Informatik e. V., 2007, pp. 227–241.

[5] A. Chacko, S. Madhu Kumar, "Big data provenance research directions," in *TENCON 2017 - 2017 IEEE Region 10 Conference*, 2017, pp. 651–656. DOI: 10.1109/TENCON.2017.8227942.

[6] D. Sacha *et al.*, "Somflow: Guided exploratory cluster analysis with self-organizing maps and analytic provenance," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 120–130, 2018. DOI: 10.1109/TVCG.2017.2744805.

[7] E. Olshannikova, A. Ometov, Y. Koucheryavy, T. Olsson, "Visualizing big data with augmented and virtual reality: Challenges and research agenda," *Journal of Big Data*, vol. 2, Oct. 2015. DOI: 10.1186/s40537-015-0031-2.

[8] M. Zbakh, M. Bakhouya, M. Essaaidi, P. Manneback, "Cloud computing and big data: Technologies and applications," *Concurrency and Computation: Practice and Experience*, vol. 30, e4517, Jun. 2018. DOI: 10.1002/cpe.4517.

[9] R. Tardío, A. Maté, J. Trujillo, "An iterative methodology for defining big data analytics architectures," *IEEE Access*, vol. 8, pp. 210 597–210 616, Jan. 2020. DOI: 10.1109/ACCESS.2020.3039455.

[10] S. Card, J. Mackinlay, "The structure of the information visualization design space," in *Proceedings of VIZ '97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium*, 1997, pp. 92–99. DOI: 10.1109/INFVIS.1997.636792.

[11] R. Hoekstra, P. Groth, "Prov-o-viz - understanding the role of activities in provenance," in *Provenance and Annotation of Data and Processes*, B. Ludäscher, B. Plale, Eds., Cham: Springer International Publishing, 2015, pp. 215–220, ISBN: 978-3-319-16462-5.

[12] W. Oliveira, L. Martins Ambrósio, R. Braga, V. Ströele, J. M. David, F. Campos, "A framework for provenance analysis and visualization," vol. 108, Jun. 2017, pp. 1592–1601. DOI: 10.1016/j.procs.2017.05.216.

[13] A. Schreiber, R. Struminksi, "Visualizing the provenance of personal data using comics," *Computers*, vol. 7, no. 1, 2018, ISSN: 2073-431X. DOI: 10.20944/preprints201712.0153.v1. [Online]. Available: https://www.mdpi.com/2073-431X/7/1/12.

[14] N. Del Rio, P. P. da Silva, "Probe-it! visualization support for provenance," in *Advances in Visual Computing*, G. Bebis *et al.*, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 732–741, ISBN: 978-3-540-76856-2.

[15] M. Kunde, H. Bergmeyer, A. Schreiber, "Requirements for a provenance visualization component," in *Provenance and Annotation of Data and Processes: Second International Provenance and Annotation Workshop, IPAW 2008, Salt Lake City, UT, USA, June 17-18, 2008. Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 241–252, ISBN: 9783540899648. [Online]. Available: https://doi.org/10.1007/978-3-540-89965-5_25.

[16] P. Chen, B. Plale, "Big data provenance analysis and visualization," *Proceedings - 2015 IEEE/ACM 15th International Symposium on Cluster, Cloud, and Grid Computing, CCGrid 2015*, pp. 797–800, Jul. 2015. DOI: 10.1109/CCGrid.2015.85.

[17] P. Missier, J. Bryans, C. Gamble, V. Curcin, "Abstracting prov provenance graphs: A validity-preserving approach," *Future Generation Computer Systems*, vol. 111, May 2020. DOI: 10.1016/j.future.2020.05.015.

[18] J. Zhao, C. Wroe, C. Goble, R. Stevens, D. Quan, R. Greenwood, "Using semantic web technologies for representing e-science provenance," Nov. 2004, pp. 92–106, ISBN: 978-3-540-23798-3. DOI: 10.1007/978-3-540-30475-3_8.

[19] B. Magagna, D. Goldfarb, P. Martin, M. Atkinson, S. Koulouzis, Z. Zhao, "Data provenance," in *Towards Interoperable Research Infrastructures for Environmental and Earth Sciences: A Reference Model Guided Approach for Common Challenges*, Z. Zhao, M. Hellström, Eds. Cham: Springer International Publishing, 2020, pp. 208–225, ISBN: 978-3-030-52829-4. DOI: 10.1007/978-3-030-52829-4_12. [Online]. Available: https://doi.org/10.1007/978-3-030-52829-4_12.

[20] M. Nacar, M. Pierce, G. Fox, "Gtlab: Grid tag libraries supporting workflows within science gateways," Nov. 2007, pp. 194–199, ISBN: 0-7695-3007-9. DOI: 10.1109/SKG.2007.136.

[21] P. Chen, B. Plale, T. Evans, "Dependency provenance in agent based modeling," Oct. 2013, pp. 180–187. DOI: 10.1109/eScience.2013.39.

[22] G. Fox *et al.*, "Real time streaming data grid applications," in Jan. 2006, pp. 253–267, ISBN: 0-387-29811-8. DOI: 10.1007/0-387-30394-4_17.

[23] S. Chaturvedi, S. Tyagi, Y. Simmhan, "Collaborative reuse of streaming dataflows in iot applications," in *2017 IEEE 13th International Conference on e-Science (e-Science)*, 2017, pp. 403–412. DOI: 10.1109/eScience.2017.54.

[24] G. Santos, J. Masó, N. Julià, X. Pons, "Geospatial queries on data collection using a common provenance model," *ISPRS International Journal of Geo-Information*, vol. 10, p. 139, Mar. 2021. DOI: 10.3390/ijgi10030139.

[25] R. Wright, "Quine: A temporal graph system for provenance storage and analysis: 7th international provenance and annotation workshop, ipaw 2018, london, uk, july 9-10, 2018, proceedings," in Sep. 2018, pp. 177–180, ISBN: 978-3-319-98378-3. DOI: 10.1007/978-3-319-98379-0_14.

[26] Y. Luo, B. Plale, S. Jensen, Y.-W. Cheah, H. Conover, "Provenance of amsr-e data from the national snow and ice data center," Ph.D. dissertation, Indiana University, 2011.

[27] K. Xu, A. Ottley, C. Walchshofer, M. Streit, R. Chang, J. Wenskovitch, "Survey on the analysis of user interactions and visualization provenance," *Computer Graphics Forum*, vol. 39, pp. 757–783, Jun. 2020. DOI: 10.1111/cgf.14035.

[28] I. M. Yazici, M. S. Aktas, "E-bilim is sureclerinde buyuk provenans verilerini gorsellestirme yontemleri," in Oct. 2020.

[29] M. Baeth, "Provenance use in social media software to develop methodologies for detection of information pollution," Ph.D. dissertation, Sep. 2019. DOI: 10.13140/RG.2.2.35217.58726.

[30] Y. Tas, M. Baeth, M. Aktas, "An approach to standalone provenance systems for big social provenance data," Aug. 2016, pp. 9–16. DOI: 10.1109/SKG.2016.010.

[31] A. Tüfek, A. Gurbuz, O. Ekuklu, M. Aktas, "Provenance collection platform for the weather research and forecasting model," Sep. 2018, pp. 17–24. DOI: 10.1109/SKG.2018.00009.

[32] M. Aktas, G. Fox, M. Pierce, "Fault tolerant high performance information services for dynamic collections of grid and web services," *Future Generation Computer Systems*, vol. 23, pp. 317–337, Mar. 2007. DOI: 10.1016/j.future.2006.05.009.

[33] M. Riveni, T.-D. Nguyen, M. Aktas, S. Dustdar, "Application of provenance in social computing: A case study," *Concurrency and Computation Practice and Experience*, Sep. 2018. DOI: 10.1002/cpe.4894.

[34] M. Aktas, S. Oh, G. Fox, M. Pierce, "Xml metadata services," Jan. 2006, pp. 31–31, ISBN: 0-7695-2673-X. DOI: 10.1109/SKG.2006.113.

[35] G. Aydin *et al.*, "Building and applying geographical information system grids," *Concurrency and Computation: Practice and Experience*, vol. 20, pp. 1653–1695, Sep. 2008. DOI: 10.1002/cpe.1312.

[36] M. Aktas *et al.*, "Iservo: Implementing the international solid earth research virtual observatory by integrating computational grid and geographical information web services," *Pure and Applied Geophysics*, vol. 163, pp. 2281–2296, Dec. 2006. DOI: 10.1007/s00024-006-0137-8.

[37] A. Sayar, M. Pierce, G. Fox, "Developing gis visualization web services for geophysical applications," Jul. 2005.

[38] B. Yildiz, K. Wu, S. Byna, A. Shoshani, "Parallel membership queries on very large scientific data sets using bitmap indexes," *Concurrency and Computation: Practice and Experience*, vol. 31, e5157, Jan. 2019. DOI: 10.1002/cpe.5157.

[39] I. M. Yazici, M. S. Aktas, M. Gokturk, "A novel approach to user involved big data provenance visualization," in *The Ninth International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA)*, 2017.

[40] P. Chen, B. Plale, Y.-W. Cheah, D. Ghoshal, S. Jensen, Y. Luo, "Visualization of network data provenance," in *2012 19th International Conference on High Performance Computing*, 2012, pp. 1–9. DOI: 10.1109/HiPC.2012.6507517.

[41] Y.-K. Suh, K. Y. Lee, "A survey of simulation provenance systems: Modeling, capturing, querying, visualization, and advanced utilization," *Hum.-Centric Comput. Inf. Sci.*, vol. 8, no. 1, Dec. 2018, ISSN: 2192-1962. DOI: 10.1186/s13673-018-0150-9. [Online]. Available: https://doi.org/10.1186/s13673-018-0150-9.

[42] A. Menin, R. Cava, C. M. D. S. Freitas, O. Corby, M. Winckler, "Towards a visual approach for representing analytical provenance in exploration processes," in *2021 25th International Conference Information Visualisation (IV)*, 2021, pp. 21–28. DOI: 10.1109/IV53921.2021.00014.

[43] I. M. Yazici, M. S. Aktas, "A usability study on data provenance visualization approaches," in *2021 15th Turkish National Software Engineering Symposium (UYMS)*, 2021, pp. 1–6. DOI: 10.1109/UYMS54260.2021.9659779.

[44] M. A. Vouk *et al.*, "Automation of network-based scientific workflows," in *Grid-Based Problem Solving Environments*, P. W. Gaffney, J. C. T. Pool, Eds., Boston, MA: Springer US, 2007, pp. 35–61, ISBN: 978-0-387-73659-4.

[45] I. Yazici, M. Aktas, "A novel visualization approach for data provenance," *Concurrency and Computation: Practice and Experience*, vol. 34, Jul. 2021. DOI: 10.1002/cpe.6523.

[46] I. M. Yazici, M. S. Aktas, "A systematic literature review on data provenance visualization," in *2th International Conference on Computing, Intelligence and Data Analytics2012 19th International Conference on High Performance Computing, ICCIDA 2022*.

[47] P. Missier, K. Belhajjame, "A prov encoding for provenance analysis using deductive rules," in *Provenance and Annotation of Data and Processes*, P. Groth, J. Frew, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 67–81, ISBN: 978-3-642-34222-6.

[48] S. H. Kaisler, F. Armour, J. A. Espinosa, W. H. Money, "Big data: Issues and challenges moving forward," *2013 46th Hawaii International Conference on System Sciences*, pp. 995–1004, 2013.

[49] B. Glavic, "Data provenance," *Foundations and Trends in Databases*, vol. 9, pp. 209–441, Apr. 2021. DOI: 10.1561/1900000068.

[50] Y. Simmhan, B. Plale, D. Gannon, "A survey of data provenance in e-science," *SIGMOD Record*, vol. 34, pp. 31–36, Sep. 2005. DOI: 10.1145/1084805.1084812.

[51] Y. G. S. Miles; “Prov model primer.” (), [Online]. Available: http://www.w3.org/TR/2013/NOTE-prov-primer-20130430/.

[52] T. L. S. S. D. McGuinness; “Prov-o: The prov ontology.” (), [Online]. Available: http://www.w3.org/TR/2013/REC-prov-o-20130430/.

[53] L. M. P. Missier; “Prov-dm: The prov data model.” (), [Online]. Available: http://www.w3.org/TR/2013/REC-prov-dm-20130430/.

[54] L. M. P. Missier; “Prov-n: The provenance notation.” (), [Online]. Available: http://www.w3.org/TR/2013/REC-prov-n-20130430/.

[55] J. C. P. M. L. Moreau; “Constraints of the prov data model.” (), [Online]. Available: http://www.w3.org/TR/2013/REC-prov-constraints-20130430/.

[56] H. H. C. T. S. Zednik; “The prov xml schema.” (), [Online]. Available: http://www.w3.org/TR/2013/NOTE-prov-xml-20130430/.

[57] G. K. P. Groth; “Provenance access and query.” (), [Online]. Available: http://www.w3.org/TR/2013/NOTE-prov-aq-20130430/.

[58] T. D. N. S. Coppens; “Prov dictionary: Modeling provenance for dictionary data structures.” (), [Online]. Available: http://www.w3.org/TR/2013/NOTE-prov-dictionary-20130430/.

[59] D. G. K. Eckert; “Dublin core to prov mapping.” (), [Online]. Available: http://www.w3.org/TR/2013/NOTE-prov-dc-20130430/.

[60] J. Cheney; “Semantics of the prov data model.” (), [Online]. Available: http://www.w3.org/TR/2013/NOTE-prov-sem-20130430.

[61] L. M. T. Lebo; “Linking across provenance bundles.” (), [Online]. Available: http://www.w3.org/TR/2013/NOTE-prov-links-20130430/.

[62] B. Kitchenham, P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, “Systematic literature reviews in software engineering-a systematic literature review,” *Information and Software Technology*, vol. 51, pp. 7–15, Jan. 2009. DOI: 10.1016/j.infsof.2008.09.009.

[63] J. Nielsen, *Usability Engineering*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, ISBN: 9780080520292.

[64] J. Brooke, “Sus: A quick and dirty usability scale,” *Usability Eval. Ind.*, vol. 189, Nov. 1995.

[65] A. Bangor, P. T. Kortum, J. T. Miller, “Determining what individual sus scores mean: Adding an adjective rating scale,” *Journal of Usability Studies archive*, vol. 4, pp. 114–123, 2009.

[66] J. Garae, R. K. L. Ko, S. Chaisiri, “Uvisp: User-centric visualization of data provenance with gestalt principles,” *2016 IEEE Trustcom/BigDataSE/ISPA*, pp. 1923–1930, 2016.

[67] Z. Cutler, K. Gadhave, A. Lex, “Trrack: A library for provenance-tracking in web-based visualizations,” Oct. 2020, pp. 116–120. DOI: 10.1109/VIS47514.2020.00030.

[68] Z. Sun, P. Yue, L. Hu, J. Gong, L. Zhang, X. Lu, "Geopwprov: Interleaving map and faceted metadata for provenance visualization and navigation," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 51, pp. 5131–5136, Nov. 2013. DOI: 10.1109/TGRS.2013.2248064.

[69] E. Wu, F. Psallidas, Z. Miao, H. Zhang, L. Rettig, "Combining design and performance in a data visualization management system," in *Conference on Innovative Data Systems Research (CIDR)*, 2017.

[70] H. Stitz, S. Luger, M. Streit, N. Gehlenborg, *Avocado: Visualization of workflow-derived data provenance for reproducible biomedical research*, Apr. 2016. DOI: 10.1101/044164.

[71] H. Stitz, S. Gratzl, H. Piringer, T. Zichner, M. Streit, "Knowledgepearls: Provenance-based visualization retrieval," *IEEE Transactions on Visualization and Computer Graphics*, vol. PP, pp. 1–1, Aug. 2018. DOI: 10.1109/TVCG.2018.2865024.

[72] M. Borkin *et al.*, "Evaluation of filesystem provenance visualization tools," *IEEE transactions on visualization and computer graphics*, vol. 19, pp. 2476–85, Dec. 2013. DOI: 10.1109/TVCG.2013.155.

[73] T. Kohwalter, T. Oliveira, J. Freire, E. Clua, L. Murta, "Prov viewer: A graph-based visualization tool for interactive exploration of provenance data," in *International Provenance and Annotation Workshop*, Jun. 2016, pp. 71–82, ISBN: 978-3-319-40592-6. DOI: 10.1007/978-3-319-40593-3_6.

[74] A. Schreiber, L. v. Kurnatowski, A. Meinecke, C. d. Boer, "An interactive dashboard for visualizing the provenance of software development processes," in *2021 Working Conference on Software Visualization (VISSOFT)*, 2021, pp. 100–104. DOI: 10.1109/VISSOFT52517.2021.00019.

[75] C. Bors, T. Gschwandtner, S. Miksch, "Capturing and visualizing provenance from data wrangling," *IEEE Computer Graphics and Applications*, vol. PP, pp. 1–1, Sep. 2019. DOI: 10.1109/MCG.2019.2941856.

[76] A. Rani, N. Goyal, S. K. Gadia, "Provenance framework for twitter data using zero-information loss graph database," ser. CODS-COMAD '21, Bangalore, India: Association for Computing Machinery, 2021, pp. 74–82, ISBN: 9781450388177. DOI: 10.1145/3430984.3431014. [Online]. Available: https://doi.org/10.1145/3430984.3431014.

[77] W. Oliveira, R. Braga, J. M. David, V. Ströele, F. Campos, G. Castro, "Visionary: A framework for analysis and visualization of provenance data," *Knowledge and Information Systems*, vol. 64, pp. 1–33, Feb. 2022. DOI: 10.1007/s10115-021-01645-6.

[78] W. de Oliveira, R. Braga, J. M. N. David, V. Stroele, F. Campos, G. Castro, "Visionary: A framework for analysis and visualization of provenance data," *Knowl. Inf. Syst.*, vol. 64, no. 2, pp. 381–413, Feb. 2022, ISSN: 0219-1377. DOI: 10.1007/s10115-021-01645-6. [Online]. Available: https://doi.org/10.1007/s10115-021-01645-6.

[79] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, S. Vassilvitskii, "Scalable k-means++," *Proc. VLDB Endow.*, vol. 5, Mar. 2012. DOI: 10.14778/2180912.2180915.

[80] L. Rokach, O. Maimon, "Clustering methods," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon, L. Rokach, Eds. Boston, MA: Springer US, 2005, pp. 321–352, ISBN: 978-0-387-25465-4. DOI: 10.1007/0-387-25465-X_15. [Online]. Available: https://doi.org/10.1007/0-387-25465-X_15.

[81] F. Nielsen, *Introduction to HPC with MPI for Data Science*. Sep. 2016, ISBN: 978-3-319-21902-8.

[82] N. Andriyanov, A. Tashlinsky, V. Dementiev, "Detailed clustering based on gaussian mixture models," in Aug. 2020, pp. 437–448, ISBN: 978-3-030-55186-5. DOI: 10.1007/978-3-030-55187-2_34.

[83] P. Chen, "Big data analytics in static and streaming provenance," Ph.D. dissertation, Indiana University, 2016.

[84] Y.-W. Cheah, B. Plale, J. Kendall-Morwick, D. Leake, L. Ramakrishnan, "A noisy 10gb provenance database," vol. 100, Aug. 2011, pp. 370–381, ISBN: 978-3-642-28114-3. DOI: 10.1007/978-3-642-28115-0_35.

[85] M. Baeth, "Provenance use in social media software to develop methodologies for detection of information pollution," Ph.D. dissertation, Sep. 2019. DOI: 10.13140/RG.2.2.35217.58726.

[86] I. M. Yazici. "Prov-o data visualization." (), [Online]. Available: https://github.com/ProvenanceVisualization/ProvOVisTool.

[87] I. M. Yazici. "Provenance data visualization research page." (), [Online]. Available: https://sites.google.com/view/data-pro-vis-research/home.

[88] I. M. Yazici. "Provenance data visualization demo." (), [Online]. Available: https://www.youtube.com/watch?v=li51oIs9Mjs&t=2s.

[89] X. Meng *et al.*, "Mllib: Machine learning in apache spark," vol. 17, no. 1, pp. 1235–1241, Jan. 2016, ISSN: 1532-4435.

[90] D. García-Gil, S. Ramírez-Gallego, S. García, F. Herrera, "A comparison on scalability for batch big data processing on apache spark and apache flink," *Big Data Analytics*, vol. 2, Mar. 2017. DOI: 10.1186/s41044-016-0020-2.

[91] T. Kodinariya, P. Makwana, "Review on determining of cluster in k-means clustering," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, pp. 90–95, Jan. 2013.

[92] I. Oakley-Girvan *et al.*, "Evaluation of a mobile device survey system for behavioral risk factors (shape): App development and usability study," *JMIR Formative Research*, vol. 3, e10246, Jan. 2019. DOI: 10.2196/10246.

[93] J. Lazar, A. Olalere, B. Wentz, "Investigating the accessibility and usability of job application web sites for blind users," *Journal of Usability Studies*, vol. 7, pp. 68–87, Jan. 2012.

[94] I. Poleszak, M. Dzieńkowski, "User experience analysis while visiting selected virtual museums," *Journal of Computer Sciences Institute*, vol. 25, pp. 384–392, Dec. 2022. DOI: 10.35784/jcsi.3057.

A
USABILITY STUDY

With a comprehensive usability study, the usability and the applicability of the studied the provenance data visualization methodologies were evaluated by the end-users as explained in Chapter 5. The usability study is conducted via a thesis research web page [87].The user study scenaros and surveys per methodology, provenance dependency querying, provenance data comparison, provenance data summarization, and streaming visualization are defined in detail to assess whether the visualization methodology meets the visualization needs in functional level after executing the task steps. In addition the general level satisfaction is assessed after executing all tasks with defined general level user surveys. In the following, first the functional level user study and the second general level study details are given.

A.1 Functional Task Level Usability Study

In order to assess the success of the studied visualization methodologies in functional level, the user tasks and the surveys are defined in detail for the visualization methodologies, backward dependency querying, Streaming visualization, provenance data comparison and the provenance data summarization.

A.1.1 Backward Provenance Dependency Querying

The backward dependency querying user study details are described for the usability evaluation in accordance with the research questions provided in Section 4.1.2 and the suggested solutions provided in 4.2.1.5. Following are the user scenarios, tasks that test subjects will complete, and user surveys that will be done after task completions.

User Scenarios and Tasks: The following test scenarios and tasks are identified for the usability test participants to execute the necessary steps before conducting the corresponding user survey.The environment preparation is given in the research web page [87] as pre-steps to be executed beforehand.

1. Selection one node from Visualization Frame, like Entity5 and click on the Get Backward Dependencies button.
2. Observation of the input nodes that cause output node selected in first step and their relations.

3. Clicking on each node in the Visualization Frame or using Show Node Properties utility defined in Utility page for further node investigation.

4. Repeating the steps by undoing the operation using the Regenerate Network option defined in Utility Page.

User Survey: A set of survey questions are defined in order to be responded by the usability test participants after the test scenario executions in order to assess whether the functional requirements are met with the visualization methodology. In the following the user survey questions are presented to be evaluated in Likert Scale.

- U1- I can focus to any visualization item to query its backward provenance.
- U2- The visualization helps me to analysis the relation between output and input nodes.
- U3- The visualization helps me to post/prior analysis between nodes.
- U4- I can do a detailed investigation for each visualized node.

A.1.2 Streaming Provenance Visualization

The streaming visualization user study details are described for the usability evaluation in accordance with the research questions provided in Section 4.1.2 and the suggested solutions provided in 4.2.1.6. Following are the user scenarios, tasks that test subjects will complete, and user surveys that will be done after task completions.

User Scenarios and Tasks: The following test scenarios and tasks are identified for the usability test participants to execute the necessary steps before conducting the corresponding user survey.The environment preparation is given in the research web page [87] as pre-steps to be executed beforehand.

1. Activation of streaming visualization by clicking the "Activate" check box as seen in fig-3.
2. Observation of the slider bar that represents time progress in visualization.
3. Observation of the visualization of data items in Visualization Frame at corresponding creation time, as seen in fig-2.
4. Playing around the data visualization with the replay buttons like forward/backward start/stop shown in fig-3.
5. Selection of Activity Based Sort option as defined in Utility Page and stopping visualization and moving the slider cursor to the start point.
6. Starting the real time visualization.
7. Observation of data items arrival in sorted order in real time, as seen in fig-2.

8. Checking the node properties individually by using Show Node Properties utility defined in Utility page by clicking on a node for further investigation of parameters.

9. Deactivation of real-time visualization for other use cases.

User Survey: A set of survey questions are defined in order to be responded by the usability test participants after the test scenario executions in order to assess whether the functional requirements are met with the visualization methodology. In the following the user survey questions are presented to be evaluated in Likert Scale.

- U1- I can trace the data items' arrival at their creation time clearly.
- U2- The visualization helps me to analysis the sequence of data items.
- U3- The visualization helps me to analysis online event tracking.
- U4- I can do a detailed investigation for each visualized node.

A.1.3 Provenance Data Summarization Visualization

The provenance data summarization user study details are described for the usability evaluation in accordance with the research questions provided in Section 4.1.2 and the suggested solutions provided in 4.2.1.3. Following are the user scenarios, tasks that test subjects will complete, and user surveys that will be done after task completions.

User Scenarios and Tasks: The following test scenarios and tasks are identified for the usability test participants to execute the necessary steps before conducting the corresponding user survey.The environment preparation is given in the research web page [87] as pre-steps to be executed beforehand.

1. Entering k cluster count for summarization, for example 4, as shown in fig-1.

2. Generation of summarization with the button "Start Clustering".

3. Observation of summarization in Visualization Frame consisting k count cluster in it.

4. Observation of the clusters' components individually by clicking on any cluster.

5. Further investigation of data items in the cluster by clicking on node items individually or by using Show Node Properties utility defined in Utility page.

6. Repeating the steps by undoing the operation using the Regenerate Network option defined in Utility Page and try another k count.

User Survey: A set of survey questions are defined in order to be responded by the usability test participants after the test scenario executions in order to assess whether the functional requirements are met with the visualization methodology. In the following the user survey questions are presented to be evaluated in Likert Scale.

- U1- The visualization helps me to summarize large-scale provenance data with data abstraction in order to provide simplicity.
- U2-I can focus on any cluster including relevant data more easily instead of having large-scale provenance data visualization.
- U3- The visualization helps me to separate data items relative to each other.
- U4- I can do a detailed investigation for each visualized node.

A.1.4 Provenance Data Comparison Visualization

The provenance data summarization user study details are described for the usability evaluation in accordance with the research questions provided in Section 4.1.2 and the suggested solutions provided in 4.2.1.4. Following are the user scenarios, tasks that test subjects will complete, and user surveys that will be done after task completions.

User Scenarios and Tasks: The following test scenarios and tasks are identified for the usability test participants to execute the necessary steps before conducting the corresponding user survey.The environment preparation is given in the research web page [87] as pre-steps to be executed beforehand. Reset Network if there is an existing network, as defined in Utility Page.

DCA Label clustering:

1. Selection of Comparison button to realize graph comparison.
2. Observation of similar-non-similar nodes in the Visualization Frame.
3. Using Show Node Properties utility defined in Utility page by clicking on a node for further investigation of parameters.
4. Resetting network as defined in Utility page.

DCA Props mode:

1. Selection of "Include all properties to comparison" in order to facilitate DCA-props mode.
2. Observation of enabling all the user configurable values in the toolbox keeping their default values.
3. Clicking of Comparison button to start comparison.

4. Observation similar and non-similar nodes visualization in the Visualization Frame.

5. Observing that different value selections lead different comparison results in the Visualization Frame. For example, it can be seen easily that the lower Threshold and Weight values the higher similarity ratio.

User Survey: A set of survey questions are defined in order to be responded by the usability test participants after the test scenario executions in order to assess whether the functional requirements are met with the visualization methodology. In the following the user survey questions are presented to be evaluated in Likert Scale.

- U1- I can apply any comparison mode easily to meet comparison needs.
- U2- Visualization helps me to understand similar and non-similar data parts efficiently.
- U3- I can configure comparison parameters via User Interface effectively to control comparison.
- U4- The visualization helps me to figure out how the different configurable values can lead different comparison results by doing subsequent comparisons effectively.
- U5- I can do a detailed investigation for each visualized node separately.

A.1.4.1 General Task Level Satisfaction Survey Questions

In order to assess the success of the studied visualization methodologies considering with the general level usability including user interface design. a general level user survey is conducted after the completion of each functional task in the previous section.

User Survey:A set of survey questions are defined in order to be responded by the usability test participants after the test scenario executions in order to assess whether the requirements are met with the visualization methodology in general. In the following the user survey questions are presented to be evaluated in Likert Scale.

- H1- The information provided within system is clear and easy to find.
- H2- It was simple to use the system.
- H3- The interface of the system is pleasant.
- H4- I can effectively complete my tasks using the system.
- H5- I can repeat the tasks and recover from mistakes easily.
- H6- The system gives feedback messages clearly.

- H7- The system gives error messages clearly.
- H8- I feel comfortable using the system.
- H9- Overall I am satisfied with the system.

PUBLICATIONS FROM THE THESIS

Papers

1. Yazici IM, Aktas MS. A novel visualization approach for data provenance. Concurrency and Computation: Practice and Experience. 2022 Apr 25;34(9):e6523.

Conference Papers

1. Yazici, Ilkay Aktas, Mehmet.A Systematic Literature Review on Data Provenance Visualization. 2th International Conference on Computing, Intelligence and Data Analytics2012 19th International Conference on High Performance Computing, ICCIDA 2022

2. Yazici IM, Aktas MS. A Usability Study on Data Provenance Visualization Approaches. In2021 15th Turkish National Software Engineering Symposium (UYMS) 2021 Nov 17 (pp. 1-6). IEEE.

3. Yazici I, Aktas M. E-Bilim Is Sureclerinde Buyuk Provenans Verilerini Gorsellestirme Yontemleri (In Turkish). International High Performance Conference (BASARIM 2020) 2020.

4. Yazici IM, Karabulut E, Aktas MS. A Data Provenance Visualization Approach. In2018 14th International Conference on Semantics, Knowledge and Grids (SKG) 2018 Sep 12 (pp. 84-91). IEEE.

5. I. M. Yazici, M. S. Aktas, M. Gokturk, "A novel approach to user involved big data provenance visualization" in The Ninth International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA), 2017.

Projects

1. Website link of the data visualization study is available at: https://sites.google.com/view/data-pro-vis-research/home

2. Youtube demo link of the data visualization study is available at: https://www.youtube.com/watch?v=li51oIs9Mjst=2s

3. Github codebase link of the data visualization study is available at: https://github.com/ProvenanceVisualization/ProvOVisTool