Research on Data Provenance Model for Multidisciplinary Collaboration

Fangyu Yu[1,2,3], Beisi Zhou[1,2,3], Tun Lu[1,2,3(✉)], and Ning Gu[1,2,3]

[1] School of Computer Science, Fudan University, Shanghai, China
{16110240028,17210240028,lutun,ninggu}@fudan.edu.cn
[2] Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China
[3] Shanghai Institute of Intelligent Electronics and Systems, Shanghai, China

Abstract. Provenance, which can be applied to assure quality, to reinforce reliability, to track fault, and to reproduce process in the end product, refers to record the lifecycle of a piece of data or thing that accounts for its generation, transformation, manipulation, and consumption, together with an explanation of how and why it got to the present place. Recently, due to its extensive applicative domains, the provenance modeling problems have brought to attention of scientific researchers significantly. In this paper, an overview of core components regarding provenance models in existing literature is presented, with a wide width from modelling methods, model comparison, and model practice, to specified issues. In addition, a collaborative model called CollabPG, was built based on the characteristics of multidisciplinary collaboration. Finally, we discussed several issues in relevance with provenance models. This paper mainly presents an overall exploration and analysis, so that potential insights could be provided for both expert and common users to select or design a provenance-based model in arbitrary applications especially multidisciplinary collaboration.

Keywords: Provenance model · Core components · Model comparison
Model practice · Multidisciplinary collaboration

1 Introduction

Different researchers interpreted definitions for provenance from different perspective. Wherein, Davidson et al. defined provenance as the data's documentation history, which includes each of conversion process's steps for the data source [1]. Herschel et al. considered provenance as "information describing the production process of some end product" [2]. Similarly, Freire et al. quoted the provenance concepts of the Oxford English Dictionary as "its history and pedigree; the source or origin of an object; a record of the ultimate derivation and passage of an item through its various owners" [3]. Ragan et al. indicated that provenance has been be used to depict the histories and origins of various types in different ways [4]. Moreover, Uri et al. depicted that provenance is a causality graph with certain nodes and edges, elucidating the process by which an object became its current state [5]. Almeida et al. mentioned that provenance, sometimes based on scientific workflows, could be utilized to preserve

Y. Sun et al. (Eds.): ChineseCSCW 2018, CCIS 917, pp. 32–49, 2019.
https://doi.org/10.1007/978-981-13-3044-5_3

particular data's execution log history as a traceable resource [6]. Allen et al. referred provenance as the record of creation, update and activities that influence a piece of data, which aids to facilitate trust in cross-organizational collaboration [7]. In this paper, a relatively common definition of provenance is proposed, which refers to record the lifecycle of a piece of data or thing that accounts for its generation, transformation, manipulation, and consumption, together with an explanation of how and why it got to the present place.

According to diverse application scenarios, provenance is mainly categorized into four categories, containing data provenance, workflow provenance, information systems provenance, and provenance meta-data [2], with a hierarchy from most general to specific ones. In the context of workflow domains, provenance possesses three types diversely: retrospective, prospective, and evolution [8–11].

In scientific researches, provenance could be employed for several purposes. For instance, scientists and engineers track provenance information to identify its contributors, occurred time, and execution process, etc., for certain data product [12]; provenance assists us to assess, maintain and improve the quality of products [13]; provenance can be used to enhance the transparency, authenticity, and integrity of a piece of data [6, 14]; In particular, scientists expend substantial effort tracking provenance data so as to ensure the repeatability and reproducibility of production process in scientific experiments [15]; It is perhaps more significant that scientists could gain insights into the chain of reasoning facilitated to discover, analyze, and explain unexpected results [16]. In a nutshell, diverse purposes provide provenance with multiple applications.

Many scholars have tackled issues with provenance across numerous domains, such as, Medical Sciences [6], Biology [17], Biomedicine [18], Genomics [19], Geography [20], and Geoinformatics [21], which were exploited in scientific workflow [22], medical records [6], financial reports [4], supply chains [23], data exploration [1], and network diagnosis [24], etc.

As illustrated in various literature, the problem of systematically modeling [25], capturing [26], storing [27], and querying [28] provenance have attracted extensive attention of scientific researchers in a wide broad of applications. In this article, we emphatically concern provenance-modeling issues in multidisciplinary collaboration applications. The aim of this article is to provide users with potential principles and sound tradeoffs while designing or choosing their peculiar provenance model. The contributions of this work are threefold. One is that we identify critical components of the provenance model and compare diverse methodologies used in them. Secondly, we conceive a collaborative model for provenance practice in multidisciplinary collaboration. Finally, we conclude certain problems existed in current model-centric provenance researches.

We organize the rest of this paper as follows. An essential outline on comparison among existing provenance-inspired models is elucidated in Sect. 2. Section 3 designs a provenance model for multidisciplinary collaboration comprehensively. Several open-ended issues on provenance models, systems, and practice are illuminated in Sect. 4. Finally, we conclude this paper with a brief conclusion of main contributions and further work.

2 Core Components of Provenance Model: An Overview

2.1 Two Classical Model Specifications

In current literature, various researchers have proposed different provenance models and relevant solutions in their respective fields. However, differences between those models make it arduous to understand the expressiveness of provenance representations, access and utilize provenance unimpededly, especially exchange information between provenance-enabled systems. Against this background, the scientific community began to emerge a consensus on provenance standardization in 2007, thus releasing and revising the open provenance model (OPM) [29] to resolve provenance-related challenges and issues. Subsequently, furtherly inspired by OPM, another conceptual model named PROV-DM [30] was endorsed by the World Wide Web consortium (W3C) in 2013, which provided well-established concepts and definitions to achieve information's interchangeable interoperability in heterogeneous contexts.

In OPM, three types and their dependencies are constituted, as shown in Fig. 1(a). Wherein, Artifact represents an immutable object during process execution, which can be expressed in physical carrier (such as device), or digital representation (such as data). Process can be considered as a range of actions to act on artifacts, and thus new artifacts may be entailed. As a contextual entity, Agent could enable, facilitate, control, and influence the execution of processes. In terms of causal relationships, one artifact, being triggered by the other, can be used or generated by a process, which may be triggered by another process, under the control of one or more agents. Similarly, PROV-DM contains core types and their relationships, forming the essence of provenance information. As depicted in Fig. 1(b), there are three element types and seven relationships. Hereinto, we consider an Entity, either real or imaginary, as something with certain fixed aspects that can be physical, digital, or conceptual. Activity performs upon or with entities during a period, and it may include generating, transforming, modifying, processing, and consuming entities. Agent is responsible for an activity's happening, the existence of an entity, or the activities of other agents.

2.2 Characteristic Comparison Among Existing Models

The OPM and PROV-DM have been currently regarded as fundamental model specifications. Despite all this, the provenance models have the variation tendency with applications and user requirements in practical usage. Instead of recreating the wheel, numerous researchers have exploited and even extended either OPM or PROV-DM to build their unified provenance model. In this section, we identify relevant studies on existing provenance models, intended to illuminate potential principle of provenance-oriented models for users, so that they could obtain insight into making informed decisions while designing or selecting a provenance model.

Review Method. In our study, an explicit strategy for literature search and selection was adopted to explore existing research works. Next, we would elaborate it gradually.

Search Strategy. First, we framed the research question (RQ) to explore focused aspects of existing provenance models, which aims at facilitating users to gain

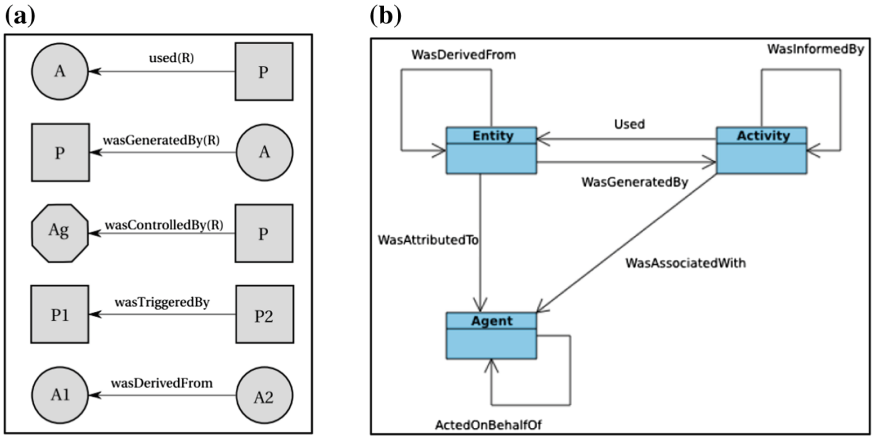

Fig. 1. (a) The OPM core composition [29], (b) The PROV-DM core composition [30]

comprehensive perspectives about the provenance-based model's principles. Further, we identified relevant studies (RS). Wherein, six databases were searched altogether in the search scope (SS): (1) IEEExplore; (2) ACM Digital Library; (3) Scopus; (4) Springer Database; (5) ScienceDirect; (6) CNKI. Based on Title, Abstract, or Keywords match, the search string (SS) was ("provenance model" OR "lineage model" OR "derivation model" OR "pedigree model") in English. Likewise, the Chinese search string was ("起源模型" OR "溯源模型" OR "世系模型"). In the filtering criteria (FC), articles pertaining to provenance model were included, and articles in the form of abstracts, summary of workshops, or systematic reviews only were excluded. We mainly surveyed outcomes of literatures between January 2014 and January 2018.

Study Selection. Initially, we obtained 608 articles from six databases via matching search strings in titles, abstracts, or keywords. Second, duplicated works (93) were excluded, with 515 articles remained. Third, we performed screenings of abstract relevance to remove 456 articles. At this point, 59 articles remained. Fourth, we reviewed the remaining articles, focusing especially on excluding those that were not related to provenance model. That is, articles (38) with no evidence of implementation, such clear statement, enforcement method, and model analysis, were removed. At this step, 21 articles remained. Finally, we used an inductive codification methodology to further analyze all full-text articles, excluding articles that were not suitable for classification. Each article used the predefined categories, including specification, type, domain, purpose, etc. As a result, 20 articles were selected totally for subsequent analyses.

Comparative Results of the Search. As depicted in Table 1, twenty provenance models are mainly surveyed. We can see from nine properties that most models [11, 31–47] utilize or extend PROV (57%) or OPM (33%), which consider their respective field features, with only rare percentages (nearly 10%) of these are built proprietarily [48] or based on other standards, such as RWS (Read-Write-Reset) [49]. In terms of element types, these are all the models pertaining to workflow (25%) or data

provenance (75%). That is to say, those models usually own specific modeling structure with certain domain, specialize on particular type of process, and apply high-level instrumentations such as structured query languages, such as SQL [50], Cypher [51], SPARQL [52], ProQL [53], etc. Particularly, in workflow-based models, all of them support some forms of retrospective provenance, whilst few of those provide related means to collect prospective (40%) and evolution (20%) provenance. However, recent literature has revealed that researchers turn towards extensions to models for capturing prospective provenance [36, 50–52], and few have proposed extensions to integrate retrospective, and evolution provenance both in design-time and run-time [11, 54]. Whereas, it is pointed that the transition is not sharp but gradient from one type to another [9].

Additionally, provenance-inspired models have been in applied various applications, under the usage of diversified purposes. Wherein, Fig. 2 indicates a wide range of domains, of which the most frequent are Biomedicine or Healthcare (20%) [32, 34, 36, 46], Security-related (15%) [33, 41, 45], Data Analysis (15%) [39, 42, 43], and Web-based (15%) [40, 47, 48], together with Domain-Agnostic (10%) and Collaboration (10%) areas [31, 36]. As shown in Fig. 3, the purpose of existing models includes Data Quality (nearly 21%), Replication (nearly 21%), Recall (nearly 18%), Data Security (nearly 15%), and Presentation (nearly 15%) dominantly.

Moreover, almost all models support compatibility (75%) and interoperability (90%), in which scalability (20%) are rarely carried. Amongst those traits, compatibility is evaluated based on its coincidence degree with these standards. For a model-enabled system with good scalability, it has the capacity to manipulate large amounts of provenance at back-end [55] and front-end [56], respectively. Interoperability is the ability to exchange provenance information within multiple systems [15, 29, 30, 54], and it is measured basically whether it is in accordance with these standards or not in this paper.

Finally, the result shows that there are three types of publications, in which large percent (75%) of them were published as conferences, 20% as journals, and one Ph.D. thesis (5%) was also covered. There is evidence that the numbers of publications remained relatively increasing (267%) from 2014 to 2017.

Be noted that, those selected models are not intended to be complete but to induce certain insight in model properties and construction for reference only, which may not cover or represent all-encompassing situations in current literature.

Discussion and Analysis. Due to space constraints, other provenance-oriented models are not enumerated in this article. However, we can probably draw three conclusions from existing literature that: (1) the majority models are core and extension of OPM or PROV-DM standards, which have been applied in a broad spectrum of domains such as biomedicine or healthcare, data analyses, web-based, security-related areas, etc., with diverse purposes of data quality, replication, recall, data security, presentation, etc. (2) almost all of available models emphasize on their compatibility, scalability, and interoperability, in which information exchange amongst systems receive attention in emerging researches. (3) existing models are mostly focused on specific discipline, whose ingredients are related to structured provenance information, with a lack of researches on unstructured collaboration especially interdisciplinary collaboration.

Table 1. Characteristic dimensions of existing provenance models (the check "√" denotes a clear statement, whilst the star "*" indicates no explicit expression in related full-text article)

No.	Name	Specification	Type				Domain	Purpose	Compatibility	Scalability	Interoperability	Publication	Year
			Workflow provenance			Data provenance							
			Prospective provenance	Retrospective provenance	Evolution provenance								
1	Interaction PM [33]	PROV				√	Service-Oriented Security	Data Security (Security Concerns, Secure Interaction)	√	*	√	Journal	2017
2	HyperFlow [49]	RWS		√			Workflow Programing	Data manipulations; Replication; Recall	*	*	√	Journal	2016
3	CRIM [36]	OPM	√	√			Biomedical Domain	Understandability; Replication; Presentation	*	√	√	Journal	2014
4	Uniform PM [11]	PROV	√	√	√		Domain-Agnostic	Replication; Presentation	√	*	√	Conference	2017
5	GeoPROV-LM [37]	PROV		√			Land Administration	Data Quality; Data Security (Access Control)	√	*	√	Conference	2015
6	SimP [38]	OPM & PROV				√	Scientific Data Management	Replication; Data Quality (Creditability)	√	*	√	Conference	2016
7	DPM [39]	OPM				√	Time Series Analysis; Data Preprocessing	Data manipulations; Replication; Recall	√	*	√	Conference	2014
8	Sensor Web-based PM [40]	PROV				√	Sensor Web	Data Quality (Usability, Reliability); Presentation	*	*	√	Conference	2017
9	Bioprov [34]	PROV				√	Biodiversity Datasets	Presentation	√	*	√	Conference	2015
10	PBAC [41]	OPM				√	Securing Provenance	Data Security (Access Authorization)	*	*	√	Conference	2015
11	CoHeal [32]	OPM				√	Healthcare Management	Data Quality	√	√	√	Conference	2017
12	Associated PM [42]	PROV				√	Statistical Analyses; Data Journalism; Scientific Research	Replication	√	*	√	Conference	2017
13	Data Point (DP) [48]	*				√	Internet of Things (IoT)	Data Quality (Trustworthiness, Dependability); Data Security	*	*	*	Conference	2017
14	Quantified Self Ontology [43]	PROV				√	Quantified Self	Data Quality (Trustworthiness); Data Security; Presentation	√	*	√	Conference	2016

(*continued*)

Table 1. (*continued*)

No.	Name	Specification	Type				Domain	Purpose	Compatibility	Scalability	Interoperability	Publication	Year
			Workflow provenance			Data provenance							
			Prospective provenance	Retrospective provenance	Evolution provenance								
15	CDPM [35]	OPM				√	Collaboration Design Process	Recall (Process Analysis)	√	*	√	Conference	2015
16	Data Supply Chain Model [44]	PROV				√	Domain-Agnostic	Recall (Process Analysis)	√	√	√	Conference	2015
17	HACCP [45]	PROV		√			Food-Safety Monitoring	Recall (Process Analysis, Process Inferring)	√	*	*	Conference	2016
18	ProvCaRe [46]	PROV				√	Biomedical Domain; Healthcare Domain	Replication; Reproducibility	√	√	√	Conference	2017
19	ROV-Model [47]	PROV				√	Social Media (Weibo)	Data Quality (Trustworthiness)	√	*	√	Journal	2017
20	CPM [31]	OPM				√	Biosciences Collaboration	Understandability; Recall (Process Analysis)	√	*	√	Phd.Thesis	2014

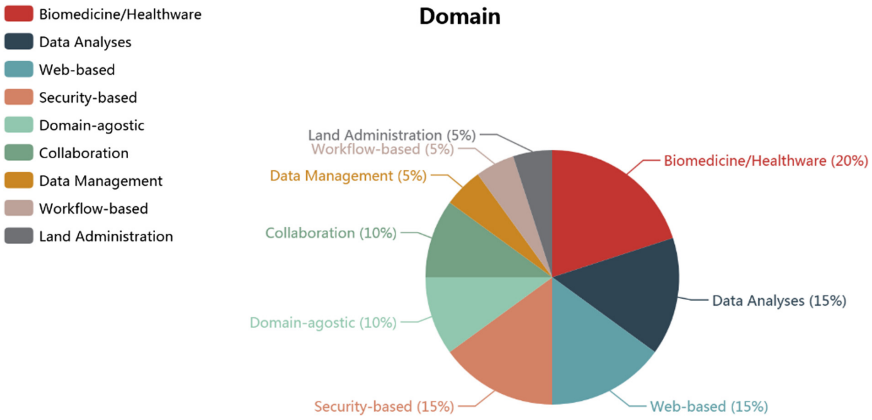

Fig. 2. Domains of existing provenance models

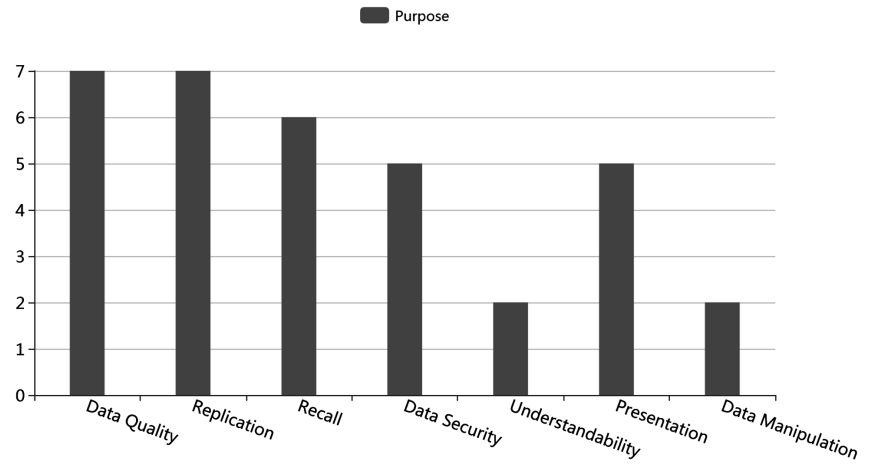

Fig. 3. Purposes of existing provenance models

3 A Collaborative Provenance Model for Multidisciplinary Applications

In the field of multidisciplinary collaboration, it is requisite that multiple researchers from different disciplines, such as physics, chemistry, computer, medicine, etc., complete creative and intellectual labor together by means of exchanging and sharing various resources. In this section, we summarize specific characteristics of multidisciplinary collaboration, and design a provenance model to record collaborative process and its associated data evolution for research collaboration across disciplines. On this basis, we also concisely evaluate the proposed model's effectiveness.

3.1 Multidisciplinary Collaboration Characteristics

Here, four features are identified by categorizing varying collaboration patterns, complex team composition, different communication schema, and dynamic collaborative process, and each of them is ever-changing in collaboration process. All those characteristics enable it challenging to design one provenance-base model that could depict the process of human interaction and data evolution wholly.

3.2 Typical Scenario

For ease of exposition in a review paper such as this, we simplified the actual scenario of multidisciplinary collaboration. As illustrated in Fig. 4, the multidisciplinary collaboration is a process of problem-solving to work together towards a common goal, via exchanging and sharing diverse resources such as hardware devices, system softwares, and information technologies among cross-disciplinary researchers, during which relevant scientific data would be generated, transformed, modified and consumed continually by those collaborators. Overall, this kind of collaboration consists two subprocesses, i.e., human interaction and data evolution, whose influence acts upon each other.

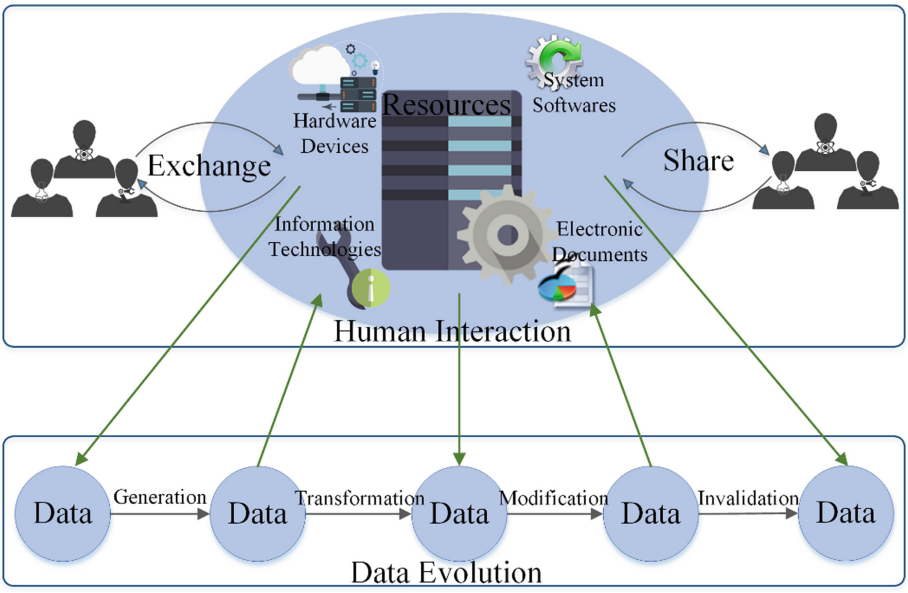

Fig. 4. The multidisciplinary collaboration scenario

3.3 A Collaborative Provenance Model: CollabPG

Based on PROV-DM [30], we further extend our collaborative provenance model, which constitutes two kinds of information: components and dependencies. In this method, we utilize a directed acyclic graph CollabPG(R, A, RE, RU, E) to collect associated provenance information, in which R, A, RE, RU are vertex sets in triple expression and E are edge sets. As shown in Fig. 5, we give an example of this model. Here, Resources (R) is expressed in yellow ovals, Activities (A) in blue rectangles, Researchers (RE) in orange pentagons, and Rules (RU) in green circles. The attributes of each element are shown in gray. In two blue cloud-patterned scopes, we can observe scientific collaboration among multidisciplinary researchers under various rules. The Black scope reveals data evolution process, including its generation, transformation, and modification.

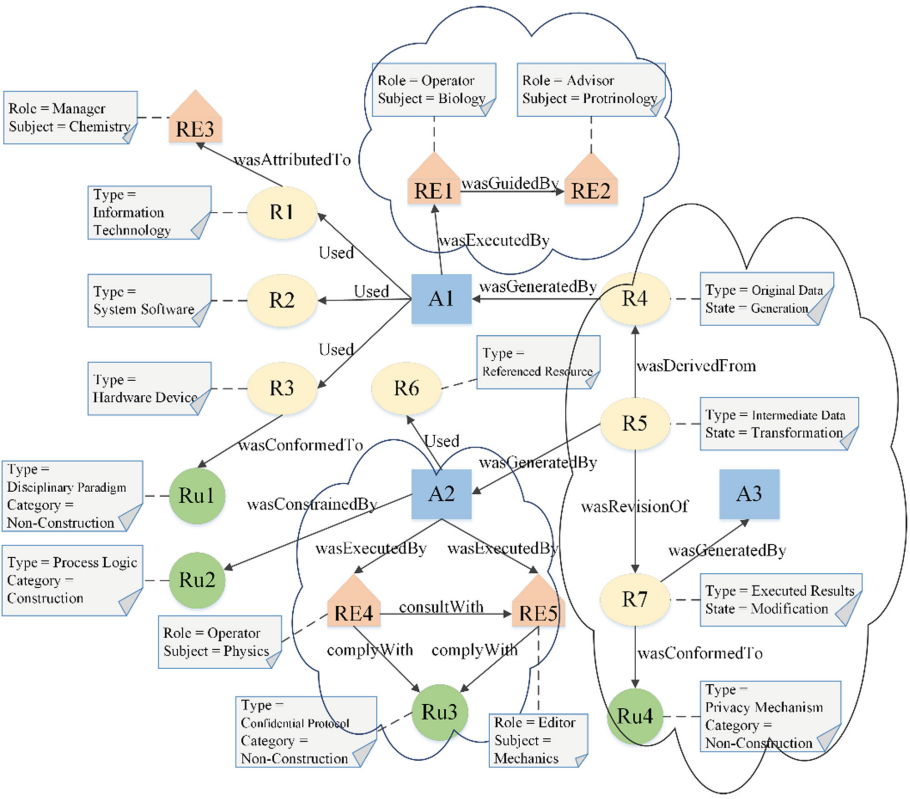

Fig. 5. An example of collaborative provenance model (Color figure online)

Components. There are four element types, including:

Resource(rid, attributes, state): denotes multiple resources that can be any physical, digital, or conceptual artifacts with certain utility values, where rid is a unique

identifier. Attributes are sets of attribute-pairs representing fixed aspects of this resource, such as the type attribute, which contains information resource (hardware devices, information technologies, system software, etc.) and scientific data (referenced resource, intermediate data, executed results, etc.) that may be collected in electronic documents. State is the resource's lifecycle phase, including generation, transformation, modification, and invalidation. The generated resource begins to be utilized, and is no longer available for use after invalidation.

Activity(aid, attributes, timeRange): refers to the collaborative activity, acting upon or with resources, which happens during a period of time. Wherein, aid is the unique identifier, and attributes are sets of attribute-pairs, such as the type attribute. Besides, the timeRange, written by [startTime, endTime], denotes the time interval (includes the beginning and end time) that an activity occurs.

Researcher (reid, attributes, subject): denotes scientific researchers responsible for the occurrence of an activity, or certain resource's existence. Wherein, reid is the unique identifier, attributes are sets of attribute-pairs such as the type attribute, which contains role, level, etc., and subject is the discipline that one researcher belongs to, such as physics, chemistry, biology, mathematics, mechanics, etc.

Rule (ruid, attributes, category): denotes sets of restriction rules that various resources, activities, or researchers have to obey. Amongst of all items, ruid is the unique identifier, and attributes are sets of attribute-pairs. The category contains structured and unstructured rules, in which the former represent process logics such as causality and concurrency. The latter are disciplinary paradigm, confidential protocol, and privacy mechanism.

Dependencies. In this model, the edge E expresses dependency relationships between above vertices. Here, sixteen dependencies are included primarily:

wasDerivedFrom(r2, r1) $\in$ R2xR1: Transforming one resource into another, i.e., R2 is transformed from R1, together with changes of certain attributes.
wasRevisionOf(r2, r1) $\in$ R2xR1: Modifying from the resource to a newest one, i.e., R2 is the revised version of R1, only minor values being updated at the same attributes.
wasGeneratedBy(r, a) $\in$ RxA: Producing one resource by an activity, i.e., Activity A generates the resource R.
Used(r, a) $\in$ RxA: Utilizing one resource by an activity, i.e., Activity A uses an existing resource R.
wasInvalidatedBy(r, a) $\in$ RxA: Invalidating the resource by an activity, i.e., Activity A invalids an existing R, due to its destruction, cessation, or expiry.
wasAttributedTo(r, re) $\in$ RxRE: Ascribing one resource with the researcher, i.e., Researcher RE is responsible for the existence of Resource R.
wasExchangedBy(a2, a1) $\in$ A2xA1: Exchanging specific resources by two activities, i.e., Activity A2 uses some resources generated by Activity A1.
wasExecutedBy(a, re) $\in$ AxRE: Executing an activity by the researcher, i.e., Researcher RE plays a role in Activity A.
dependOn(re2, re1) $\in$ RE2xRE1: Researcher RE2's outcome depends on importing contributions of RE1.

consultWith(re2, re1) ∈ RE2xRE1: Researcher RE2 carries out activities together with RE1 via joint supervision, consultation, and decision-making.
wasGuidedBy(re2, re1) ∈ RE2xRE1: Researcher RE2 directly acts on activities under the guidance of RE1.
wasConformedTo(r, ru) ∈ RxRU: Conforming one resource to certain rule, i.e., Resource R conforms to the Rule RU, such as disciplinary paradigm.
wasConstrainedBy(a, ru) ∈ AxRU: Restricting an activity to certain rule, i.e., Activity A was constrained by Rule RU, such process logic.
complyWith(re, ru) ∈ RExRU: Complying researcher's behavior with certain rule, i.e., Researcher RE complies with Rule RU, such as confidential protocol.
exclusiveWith(ru2, ru1) ∈ RU2xRU1: Rule RU2 and RU1 is mutually exclusive.
Precede(ru2, ru1) ∈ RU2xRU1: Rule RU2 have priority over RU1, whatever they are exclusive or not.

3.4 Evaluation of the CollabPG Model

Here, we mainly focus three evaluation criteria on our model, which contains compatibility, scalability, and interoperability.

The PROV-DM [30] defines some conceptual standards, such as information collection, storage methods, and query technologies, aiming to achieve the goal of exchanging information between heterogeneous systems. The proposed CollabPG model has the compatibility with it. Wherein, the resource, activity, and researcher have similar functionality to entity, activity, and agents in PROV-DM. Considering such factors as privacy, sensitivity, and control-flow of provenance information, we add the element of rule and related dependencies in our model. The relationships, such as wasExchangedBy and wasExecutedBy, correspond to wasInformedBy and wasAssociatedWith, while dependOn, consultWith, and wasGuidedBy can be viewed as extends of actedOnBehalfOf. Through the analysis above, we can build a mapping from the collaborative provenance model to the PROV-DM, so that our model ensures its compatibility, which supports exchanging information with other provenance-enabled models. Specially, collaboration characteristics have been reflected explicitly in our model, whose comparison with PROV-DM is shown concretely in Table 2.

Besides, it can be observed that our model has the interoperability to support exchange information amongst multiple systems, due to its accordance with the PROV-DM standard. Moreover, we would pursue good scalability in subsequent model-based system design as well.

4 Challenges and Opportunities on Provenance Model in Multidisciplinary Collaboration

After surveying the state of the art, this section would concern specific research issues on the balance between models, systems, and practice in provenance exploration of multidisciplinary collaboration. We would introduce each of them separately.

Table 2. Comparison of collaborative provenance model with PROV-DM

		CollabPG	PROV-DM
Ingredients	Basic elements	Resource	Entity
		Activity	Activity
		Researcher	Agent
	Extended element	Rule	–
Dependencies	Data-evolution-related relationships	wasDerivedFrom	wasDerivedFrom
		wasRevisionOf	wasRevisionOf
		wasGeneratedBy	wasGeneratedBy
		Used	Used
		wasInvalidatedBy	wasInvalidatedBy
	Basic relationships	wasAttributedTo	wasAttributedTo
		wasExchangedBy	wasInformedBy
		wasExecutedBy	wasAssociatedWith
	Collaborative relationships	dependOn	actedOnBehalfOf
		consultWith	actedOnBehalfOf
		wasGuidedBy	actedOnBehalfOf
	Rule-based relationships	wasConformedTo	–
		wasConstrainedBy	–
		complyWith	–
		exclusiveWith	–
		Precede	–

Trade-Off Between Core Principles and Extension Requirements for Provenance-Bound Models. Concerning the core criteria of a provenance model to be quality-guaranteed, several researchers have summarized related criteria for evaluating the quality of models. Examples include Completeness, Correctness, Clarity, Consistency, Simplicity and Comprehensibility. That is, the model should contain all ingredients of the domain that are relevant, conformed to the syntax of modeling language together with authentic and correct information. Moreover, the statements in the model are not uncontested, contradictory, and redundant. Lastly, the model should be effortless to be understood by its users and developers. Meanwhile, it may be inevitable to adjust provenance via extending original components of models to satisfy specific needs in practical applications. Under this circumstance, qualified requirements such as compatibility, scalability, and interoperability, could enhance the capacities of models. Therefore, it is anticipant that core and extension in provenance models are considered comprehensively in the future approach.

Trade-Off Between General Models and Specific Systems. On one hand, a desired model is versatilely used, domain-agnostic, loosely-coupled with systems, and supports interoperability and interchange among systems as well. On the other hand, no any model is likely to be self-contained and represent all provenance-inspired systems. Sometimes, the representativeness of one model is more imperative than its completeness. In practical exploration, the model should be integrated with specific

application system. For instance, there is a correlation between provenance models and capture systems, in which the information granularity of models mobilizes diverse grained-level systems to be adopted. However, we have to take integration efforts, provenance granularity, false positives, and analysis scope into count in terms of choosing appropriate capture methods and systems [26]. Specifically, the granularity of capture encounters provenance costs, i.e., fine-grained capturing approaches aggravate the issues of information overload, performance influence, and memory workload. Therefore, we should pertinently select coarse-grained, fine-grained, or hybrid-enabled systems based on actual models and scenarios.

Trade-Off Between Privacy and Utility of Information in Provenance Model. Several studies have indicated that attentions with provenance-centric disclosure are linked to issues of security and privacy concerns [16, 22], due to part of provenance's sensitivity and confidentiality, particularly for individual interaction from different disciplines in collaborative environments. When it comes to security, researchers exploit diverse access control strategies, such as authentication, authorization, and sandboxing, aiming to pinpoint which view of provenance that particular users can access. As elucidated in existing literature, customized techniques, including sanitization, abstraction, obscuring, and redaction, are employed to render an abstracted overview of provenance by omitting sensitive pieces of its information. However, those pruning methods inevitably yields varying degree of utility loss for provenance usage, which may pose possibly undesirable side-effects while exploring provenance details. As a consequence, it remains to be explored to consider double-side factors about balancing an appropriate threshold of confidentiality protection and utility preservation in provenance information. More specifically, we could reveal partial provenance to targeted users varying with their ownership roles, trust levels, and access privileges. At the same time, fractional information could be concealed according to its sensitive attributes, privacy requirements, and application propensity.

In the domain of multidisciplinary collaboration, challenges mentioned above involve only some issues of models, whose proposals may be applicable to arbitrary applications as well. In provenance practice, one important point is that choice of what model-based solution is most appropriate depends on different needs.

5 Conclusion and Future Work

In this paper, we revealed underlying overview that constitute core components of provenance model, such as, model specification, characteristic comparison, and model analysis. We conceived a collaborative model with multi-faceted factors in multidisciplinary applications, designed to depict cross-disciplinary scientists' collaboration process through exchanging and sharing diverse resources, together with its associated provenance data evolution. We summarized fundamental issues in existing provenance models to facilitate the understanding of model dimensions and construction. A recapitulative research in this article was designed to facilitate to make reasonable decisions about which model-based provenance solution to choose for both domain experts and common users in interdisciplinary applications.

In the future research, we intend to further explore dependency path calculations, tracking mechanisms, storage methods, query technologies, and access visualization of provenance applied in multidisciplinary applications, combined with their collaborative characteristics and attributes.

Acknowledgment. This work was supported by the Joint Fund of National Natural Science Foundation of China and the China Academy of Engineering Physics (NSAF) under Grant No. U1630115, and the National Key Research and Development Program of China under Grant No. 2018YFC0381402.

References

1. Davidson, S.B., Freire, J.: Provenance and scientific workflows: challenges and opportunities. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 1345–1350. ACM, London (2008)
2. Herschel, M., Hlawatsch, M.: Provenance: on and behind the screens. In: Proceedings of the 2016 International Conference on Management of Data, pp. 2213–2217. ACM, London (2016)
3. Freire, J., Koop, D., Santos, E., Silva, C.T.: Provenance for computational tasks: a survey. Comput. Sci. Eng. **10**(3), 11–21 (2008)
4. Ragan, E.D., Endert, A., Sanyal, J., Chen, J.: Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes. IEEE Trans. Vis. Comput. Graph. **22**(1), 31–40 (2016)
5. Braun, U., Shinnar, A., Seltzer, M.: Securing provenance. In: Proceedings of the 3rd Conference on Hot Topics in Security, p. 4. USENIX Association (2008)
6. Almeida, F.N., Tunes, G., da Costa, J.C.B., Sabino, E.C., Junior, A.M., Ferreira, J.E.: A provenance model based on declarative specifications for intensive data analyses in hemotherapy information systems. Future Gener. Comput. Syst. **59**, 105–113 (2016)
7. Allen, M.D., Chapman, A., Seligman, L., Blaustein B.: Provenance for collaboration: detecting suspicious behaviors and assessing trust in information. In: International Conference on Collaborative Computing: Networking, Applications and Worksharing, pp. 342–351. IEEE, Washington (2012)
8. Zafar, F., et al.: Trustworthy data: a survey, taxonomy and future trends of secure provenance schemes. J. Netw. Comput. Appl. **94**, 50–68 (2017)
9. Herschel, M., Diestelkämper, R., Lahmar, H.B.: A survey on provenance: what for? What form? What from? VLDB J. **5**, 1–26 (2017)
10. Pimentel, J.F., Freire, J., Braganholo, V., Murta, L.: Tracking and analyzing the evolution of provenance from scripts. International Provenance and Annotation Workshop (2016)
11. Duan, X., et al.: Linking design-time and run-time: a graph-based uniform workflow provenance model. In: IEEE International Conference on Web Services, pp. 97–105. IEEE, Washington (2017)
12. Cheney, J., Chiticariu, L., Tan, W.C.: Provenance in databases: why, how, and where. Found Trends Databases **1**(4), 379–474 (2009)
13. Ross, S.: Digital preservation, archival science and methodological foundations for digital libraries. New Rev. Inf. Netw. **17**(1), 43–68 (2012)
14. Boose, E.R., Ellison, A.M., Osterweil, L.J., Clarke, L.A., Podorozhny, R., Hadley, J.L., Wise, A.E., Foster, D.R.: Ensuring reliable datasets for environmental models and forecasts. Ecol. Inform. **2**(3), 237–247 (2007)

15. Groth, P., Moreau, L.: PROV-overview: an overview of the PROV family of documents (2013)
16. Bachour, K., Wetzel, R., Flintham, M., Huynh, T.D., Rodden, T., Moreau, L.: Provenance for the people: an HCI perspective on the W3C PROV standard through an online game. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 2437–2446. ACM, London (2015)
17. Zhao, J., Miles, A., Klyne, G., Shotton, D.: Provenance and linked data in biological data webs. Brief. Bioinform. **10**(2), 139–152 (2008)
18. Masseroli, M., Canakoglu, A., Ceri, S.: Integration and querying of genomic and proteomic semantic annotations for biomedical knowledge extraction. IEEE/ACM Trans. Comput. Biol. Bioinf. **13**(2), 209–219 (2016)
19. Ocaña, K.A., Silva, V., De Oliveira, D., Mattoso, M.: Data analytics in bioinformatics: data science in practice for genomics analysis workflows. In: IEEE International Conference on e-Science, pp. 322–331. IEEE, Washington (2015)
20. Zhao, H., Zhang, S., Zhang, Z.: Relationship between multi-element composition in tea leaves and in provenance soils for geographical traceability. Food Control **76**, 82–87 (2015)
21. Yue, P., He, L.: Geospatial data provenance in cyberinfrastructure. In: 2009 17th International Conference on Geoinformatics, pp. 1–4. IEEE, Washington (2009)
22. Holten Møller, N.L., Bjørn, P., Villumsen, J.C., Hancock, T.C.H., Aritake, T., Tani, S.: Data tracking in search of workflows. In: The ACM Conference on Computer-Supported Cooperative Work and Social Computing. ACM, New York (2017)
23. Li, P., Wu, T.Y., Li, X.M., Luo, H., Obaidat, M.S.: Constructing data supply chain based on layered PROV. J. Supercomput. **73**(4), 1509–1531 (2016)
24. Chen, A., Wu, Y., Haeberlen, A., Zhou, W., Loo, B.T.: The good, the bad, and the differences: better network diagnostics with differential provenance. In: Conference on ACM SIGCOMM 2016 Conference, pp. 115–128. ACM, New York (2016)
25. Bowers, S., McPhillips, T., Ludäscher, B., Cohen, S., Davidson, Susan B.: A model for user-oriented data provenance in pipelined scientific workflows. In: Moreau, L., Foster, I. (eds.) IPAW 2006. LNCS, vol. 4145, pp. 133–147. Springer, Heidelberg (2006). https://doi.org/10.1007/11890850_15
26. Stamatogiannakis, M., et al.: Trade-offs in automatic provenance capture. In: Mattoso, M., Glavic, B. (eds.) IPAW 2016. LNCS, vol. 9672, pp. 29–41. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40593-3_3
27. https://www.w3.org/TR/2013/REC-prov-o-20130430/
28. Wylot, M., Cudremauroux, P., Hauswirth, M., Groth, P.: Storing, tracking, and querying provenance in linked data. IEEE Trans. Knowl. Data Eng. **29**, 1751–1764 (2017)
29. Moreau, L., et al.: The open provenance model core specification (v1.1). Fut. Gener. Comput. Syst. **27**(6), 743–756 (2011)
30. Missier, P., Belhajjame, K., Cheney, J.: The W3C PROV family of specifications for modelling provenance metadata. In: Proceedings of EDBT, pp. 773–776 (2013)
31. Huang, X.: Research on biology collaboration: scientific software sharing, selection and recommendation. Ph.D. thesis, Fudan University (2014) (in Chinese)
32. Sun, Y., Lu, T., Gu, N.: A method of electronic health data quality assessment: enabling data provenance. In: Proceedings of CSCWD 2017. IEEE, Washington, pp. 233–238 (2017)
33. Hasan, R., Khan, R.: Unified authentication factors and fuzzy service access using interaction provenance. Comput. Secur. **67**, 211–231 (2017)
34. Amanqui, F.K., et al.: A model of provenance applied to biodiversity datasets. In: 2016 IEEE 25th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), pp. 235–240. IEEE, Washington (2016)

35. Sun, X., Gao, X., Kang, H., Li, C.: A data provenance model for collaboration design process. In: International Conference on Information Sciences, Machinery, Materials and Energy (2015)
36. Curcin, V., Miles, S., Danger, R., Chen, Y., Bache, R., Taweel, A.: Implementing interoperable provenance in biomedical research. Future Gener. Comput. Syst. **34**, 1–16 (2014)
37. Sadiq, M.A., West, G., McMeekin, D.A., Arnold, L., Moncrieff, S.: Provenance ontology model for land administration spatial data supply chains. In: International Conference on Innovations in Information Technology, pp. 184–189. IEEE, Washington (2016)
38. Jabal A A., Bertino E.: SimP: secure interoperable multi-granular provenance framework. In: International Conference on E-Science, pp. 270–275. IEEE (2017)
39. De Souza, L., Vaz, M.S.M.G., Sunye, M.S.: Modular development of ontologies for provenance in detrending time series. In: International Conference on Information Technology: New Generations, pp. 567–572. IEEE Computer Society, Washington (2014)
40. Jiang, L., Kuhn, W., Yue, P.: An interoperable approach for Sensor Web provenance. In: International Conference on Agro-Geoinformatics, pp. 1–6 (2017)
41. Mohy, N.N., Mokhtar, H.M.O., El-Sharkawi, M.E.: Delegation enabled provenance-based access control model. In: Science and Information Conference, pp. 1374–1379. IEEE, Washington (2015)
42. Trinh, T.D., et al.: Linked data processing provenance: towards transparent and reusable linked data integration. In: The International Conference, pp. 88–96 (2017)
43. Schreiber, A.: A provenance model for quantified self data. In: International Conference on Human–Computer Interaction (2016)
44. Lan, J., Liu, X., Luo, H., Li, P.: Study of constructing data supply chain based on PROV. In: Wang, Yu., Xiong, H., Argamon, S., Li, X., Li, J. (eds.) BigCom 2015. LNCS, vol. 9196, pp. 69–78. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22047-5_6
45. Markovic, M., Edwards, P., Kollingbaum, M., Rowe, A.: Modelling provenance of sensor data for food safety compliance checking. In: Mattoso, M., Glavic, B. (eds.) IPAW 2016. LNCS, vol. 9672, pp. 134–145. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40593-3_11
46. Valdez, J., Rueschman, M., Kim, M., Arabyarmohammadi, S., Redline, S., Sahoo, S.S.: An extensible ontology modeling approach using post coordinated expressions for semantic provenance in biomedical research. In: Panetto, H., et al. (eds.) On the Move to Meaningful Internet Systems, OTM 2017 Conferences, OTM 2017. LNCS, vol. 10574. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69459-7_23
47. Zhang, Z., Dong, H., Tan, C., Yi, Y.: Evaluation of Weibo credibility based on data provenance. In: Application Research of Computers (2017) **(in Chinese)**
48. Olufowobi, H., Engel, R., Baracaldo, N., Bathen, Luis Angel D., Tata, S., Ludwig, H.: Data provenance model for Internet of Things (IoT) systems. In: Drira, K., et al. (eds.) ICSOC 2016. LNCS, vol. 10380, pp. 85–91. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68136-8_8
49. Balis, B.: HyperFlow: a model of computation, programming approach and enactment engine for complex distributed workflows. Future Gener. Comput. Syst. **55**, 147–162 (2016)
50. Barga, R.S., Digiampietri, L.A.: Automatic capture and efficient storage of eScience experiment provenance. Concurr. Comput. Pract. Exp. **20**(5), 419–429 (2008)
51. https://neo4j.com/developer/cypher-query-language/
52. https://www.w3.org/TR/rdf-sparql-query/
53. Karvounarakis, G., Ives, Z.G., Tannen, V.: Querying data provenance. In: ACM Conference on the Management of Data (SIGMOD), pp. 951–962 (2010)

54. Bowers, S., McPhillips, T., Riddle, S., Anand, M.K., Ludäscher, B.: Kepler/pPOD: scientific workflow and provenance support for assembling the tree of life. In: Freire, J., Koop, D., Moreau, L. (eds.) IPAW 2008. LNCS, vol. 5272, pp. 70–77. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89965-5_9
55. Akoush, S., Sohan, R., Hopper, A.: HadoopProv: towards provenance as a first class citizen in MapReduce. In: Usenix Workshop on the Theory and Practice of Provenance. USENIX Association (2013)
56. Deutch, D., Gilad, A., Moskovitch, Y.: selP: selective tracking and presentation of data provenance. In: International Conference on Data Engineering, pp. 1484–1487. IEEE, Washington (2015)