A Systematic Review of Provenance Systems

Supplementary material

Beatriz Pérez, Julio Rubio, and Carlos Sáenz-Adán

1 Introduction

In this document we show some aspects regarding the performance of the systematic review entitled "A systematic review of provenance systems", which has been undertaken based on the original guidelines as described in [1], [2], [3]. The systematic review has been performed in two review processes (we will refer to them as *first review process* and *second review process*, respectively). More specifically, the *first review process*, which was carried out from September 2015 to June 2016, included research studies published up to and including December 2015. This first review process was renewed by performing a *second review process*, which took place between July and August 2017, and which covered research studies published from January 2016 to July 2017. Thus, the overall systematic review covers studies published up to and including July 2017.

Taking this into account, in this document we address three main issues. First, we describe in detail the methodology followed to perform the first review process. Secondly, we present the main remarks regarding the performance of the second review process. Finally, we discuss the study limitations, analyzing the threats to validity arising from the procedures we followed to perform the systematic review.

2 The First Review Process

As explained previously, the study of provenance systems was undertaken as a systematic review based on the original guidelines as described in [1], [2], [3]. More specifically, the process we followed in the first review process is depicted in Figure 1, which represents the planning, conducting, and reporting stages we considered.

Firstly, the *planning* stage mainly dealt with identifying the need for the review (see *Phase 1* in Figure 1), and establishing a review protocol (see *Phase 2* in Figure 1). Regarding the interest for the review, we firstly performed preliminary searches aimed not only at identifying existing systematic reviews or surveys, but also assessing the volume of potentially relevant studies. This initial informal search confirmed that there were a substantial number of papers on the topic (none of them performed as a systematic review), and that a systematic review would be appropriate. Concerning a review protocol, it states an accurate and reliable methodology to perform the review and mainly specifies: research questions to be addressed, inclusion and exclusion criteria, search and selection strategy, quality assessment and data extraction, and the data synthesis.

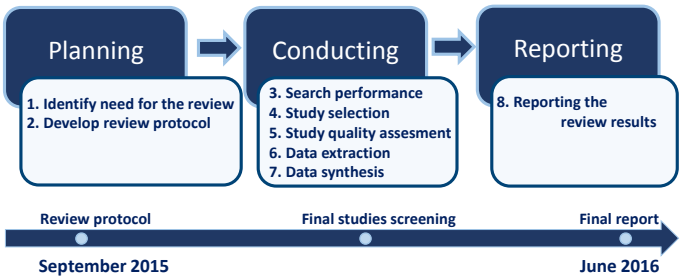

Fig. 1. First review process of the systematic review

Secondly, after planning the review, we started the *conducting* stage by following the protocol as it had been defined (see Figure 1). Firstly, we performed the search strategy. As we will explain later, in this phase we specially benefited from preliminary searches which specially help us to identify existing studies and assessing the volume of potentially relevant studies (see *Phase 3*). Then, we went on with the next phase (see *Phase 4* in Figure 1) by identifying the relevant studies taking into account the research questions. In this step we also consider the inclusion and exclusion criteria of each study as stated in our protocol.

Later, we performed the quality assessment to analyze and assess the primary studies (see *Phase 5*). *Phase 6* involved filling the data extraction forms defined to accurately record the information obtained from the primary studies in order to answer the research questions. In the last final step we performed the data synthesis by collating and summarizing the results of the included primary studies (see *Phase 7*). It is worth noting that, as we will present later in more detail, due to our specific research questions, we performed these two phases 6 and 7 twice. Finally, the *reporting* stage took place (see *Phase 8* in Figure 1).

Next, we document in more detail the different phases considered to conduct our protocol, regarding the *planning* stage and part of the *conducting* stage. More specifically, the results obtained from the data synthesis of the studies selected in this review process are presented in the original paper [4].

• *B. Pérez, J. Rubio, and C. Sáenz-Adán were with the Department of Mathematics and Computer Science University of La Rioja, E-26004 – La Rioja, Spain.*

2.1 Research Questions

The research questions addressed by this study are the following:

RQ1 What are the different aspects to take into account in provenance systems?
RQ2 What are the proposed techniques to address such different aspects?
RQ3 Taking into account the most representative provenance systems within the provenance field, how these systems have addressed the identified aspects?

Regarding question *RQ1*, we aimed at identifying the different features or aspects in which provenance systems could be categorized, establishing a taxonomy of features. From this study, with question *RQ2*, we also wanted to know what are the proposed techniques or solutions that exist so far to address such different features. To this aim, our intention consisted in identifying the techniques, methodologies, or approaches available in the literature. As for *RQ3*, we aimed at looking for evidences of these existing provenance systems which have somehow an impact or repercussion within the provenance community, assessing their characteristics and quality, in terms of our taxonomy.

2.2 Inclusion and Exclusion Criteria

In order to make sure that the studies included in the review were clearly related to the research topic, we defined detailed general guidelines for inclusion and exclusion criteria. In particular, the scope of the systematic review is limited to the literature that: (i) presents or describes solutions for research in provenance systems within the computer science context, and/or (ii) surveys or analyzes the previous type of studies. In particular, given the knowledge provided by these latter type of literature, we decided to identify the published surveys and reviewing papers on the topic as potentially relevant for the study. We did not impose any restrictions on a specific domain of application. Moreover, this first review process included research studies published up to and including December 2015.

On the other hand, we excluded pure discussion or opinion papers, tutorials and any study that tackles provenance in a context other than the computer science field. We also exclude any study reported in a language other than English.

2.3 Search and Selection Strategy

The identification of as much literature relevant to the research questions as possible using an unbiased search strategy constitutes a necessary and crucial step in any systematic review. Different approaches are usually taken to perform the search strategy. In particular, guidelines [1], [2], [3] emphasize to carry out the literature search by performing an automatic search through web search engines provided by electronic databases. This strategy is also suggested to be supported by the use of other sources of evidence including journals, grey literature (i.e. technical reports, work in progress) and conference proceedings, performing if needed a manual search. In practice, systematic reviews performed in the specific field of software engineering have used different combinations of the different alternatives employing automated or manual search, alone or combined [5].

At the time we began the first review, we had a limited knowledge of provenance research and related set of venues (journals or workshops) where we could conduct our search. We had two interesting surveys obtained from informal searches, more specifically, [6] and [7] (neither of them explained the way in which they identified the analyzed studies). Thus, we initially decided to make a first attempt establishing and performing a plan based on an automatic search to identify the primary studies.

2.3.1 Automatic Search Process

Our first step towards performing an automatic search was to define the scope of the study based on the PICO template [8]. Following the PICO template, we identified the scope of the study as follows:

- *Population*: The published scientific literature reporting provenance systems and proposals.
- *Intervention*: Any study or publication presenting the characteristics of a provenance system and provenance proposals.
- *Comparison*: Different key aspects each provenance system is desired to have and the proposed techniques or solutions given for addressing them.
- *Outcome*: The completeness of the coverage of such key aspects and proposed techniques.

We started our systematic electronic search by identifying a list of keywords and search terms following a three–steps process (see Table 1). Due to the relevance of the selection of terms for the quality of results, and considering the scope of our study, we selected general terms with the aim of confirming that most of the research papers were included in the study. More specifically, in order to be as unbiased as possible, we chose a set of general keywords classified into three different groups: *provenance*, *aspect* and *technique* synonym concepts (see Table 1). Taking this into account, firstly, we started from the set of terms *provenance*, *technique* and *aspect* (see step 1 in Table 1). Secondly, we augmented this list with associated terms and synonyms, obtaining a more complete list (see step 2 in Table 1). With these keywords, we conducted a trial search that identified a new keyword related to a term usually refereed to provenance included in different proposals unknown to us at the beginning of the review (see step 3 in Table 1).

TABLE 1
Keywords used in our search

Step	Provenance conc.	Aspect conc.	Technique conc.
1	Provenance	Aspect	Technique
2		Feature Characteristic Property Issue Classify	Procedure Methodology Solution Approach Method
3	Lineage		

Based on the selected keywords and after some pilot testing, we constructed a final abstract search string containing

all the relevant keywords chosen for the search, which is the following:

```
(Provenance OR Lineage) AND
(Aspect OR Feature OR Characteristic OR Property
OR Issue OR Classify) AND
(Technique OR Procedure OR Methodology OR Solution
OR Approach OR Method)
```

To carry out the search, we performed a review through systematic reviews performed in Software Engineering (SE) which showed several choices for electronic databases (DBs), such as IEEE Xplore, the ACM Digital Library, Computer Database, ScienceDirect, Scopus, ISI Web of Science, Springer Link and Google Scholar. Taking this into account, we decided to select databases with good coverage, reputation, advanced features to perform the search and exportability (specially as BibTeX format). Firstly, we selected several electronic databases considering the following as most relevant: IEEE Xplore, the ACM Digital Library, ScienceDirect, and Scopus. Other DBs such as ISI Web of Science did not allow us to search on the full text, and Google Scholar and SpringerLink did not provide a flexible formulation of search strings with unlimited clauses.

Based on our abstract search string, different strings were derived on each selected database engine, taking into account their search particularities considered in all databases the advanced search option (in fact, we came across several problems related to the particularities of search of each DB). Some of the different searches we performed are shown in Table 2, where we have specified: (1) the electronic database in which the search took place, (2) the fields in which the search took place (mainly title, abstract and keywords, when possible, and full text), (3) the area of search, and (4) the number of papers resulting from the search. We do not show each detailed search strings due to space reasons. We note that such a number refers to papers without considering any selection criteria other than the range of years (from 1970 to last 2015) and the language of publication (English). Regarding the search location in the papers, we started our search looking mainly in titles, abstracts, and keywords (see results labeled with *a* in Table 2). A quick look at the results shows us that some of the previously identified known set of papers were not included, which lead us to search on the full text (see results labeled with *b* in Table 2), obtaining, as it was expected, an unmanageable amount of results. Considering previous systematic reviews published in the literature in Software Engineering, we realized that the most commonly used subjects are 'computer science' or 'software engineering'. For this reason, to limit the amount of papers, we try to refine the search by looking by the 'computer science' field when possible (only the search in 'computer science' was available in ScienceDirect and Scopus, where no field filter was possible to be applied in IEEE Xplore and ACM), obtaining the results labeled with *c* in Table 2. As a result, the research found in these electronic databases yielded a total of 44690 unmanageable results (considering the results within the 'computer science' field in ScienceDirect and Scopus, plus the results within all fields in IEEE Xplore and ACM). Focusing only on the results within the 'computer science' field, we made an experiment taking into account only the results obtained from ScienceDirect and Scopus (labeled with *c* in Table 2). We used the *export citation* feature provided by ScienceDirect and Scopus, and we obtain the results in BibTex format. Later, we joined the two BibTex files obtaining a file with a total of 2001+4243= 6244 references. We ruled out duplicates and excluded some results which clearly not related to the research focus (these which correspond to conference presentations, publisher summaries, bibliographies, presentation contents, etc.), obtaining a total of 5764 possible candidates, only considering results from ScienceDirect and Scopus restricted to the 'computer science' field.

TABLE 2
Initial searches

DB	id	Place of search	Area	Num.
Science Direct	a	Title, keywords and abstract	All sciences	1.849
	b	Full text	All sciences	130.462
	c	Full text	Computer sciences	2.001
Scopus	a	Title, keywords and abstract	All sciences	8.715
	b	Full text	All sciences	155.083
	c	Full text	Computer sciences	4.243
IEEExplore	a	abstract (zero results in Title and keywords, performed in separate searches)	All sciences	155
	b	Full text	All sciences	5.891
ACM	a	Title, abstract (separate searches)	All sciences	0
	b	Full text	All sciences	32.555

Thus, the results obtained by our automated search turned out to show that there is a huge number of research papers on the topic, even when we limited the search to the 'computer science' field when it was possible. Such a wide number of results could be particularly justified by existing research on the topic of performing systematic reviews within the specific software engineering domain [9]. More specifically, in [9] authors point out the fact that (1) current software engineering digital databases do not provide good support for the identification of relevant research in systematic literature reviews, and (2) in contrast to other researchers such as medical ones, software engineering researchers need to perform resource-dependent searches.

Taking this into account, and checking practices of other researchers looking at research trends [3], [10], we thought about the possibility of performing a manual search process of a specific set of venues, but as stated in [3], it often implies missing relevant studies from excluded venues. For this reason, we finally decided to perform our systematic review by following an alternative search strategy that basically combined an automatic search with a manual search.

2.3.2 Combining an Automatic search and a Manual Search

In our second attempt, the search and selection process was performed in three stages, as outlined in Figures 2-4 respectively, and which are described in more detail in the following subsections. The overall process was supported

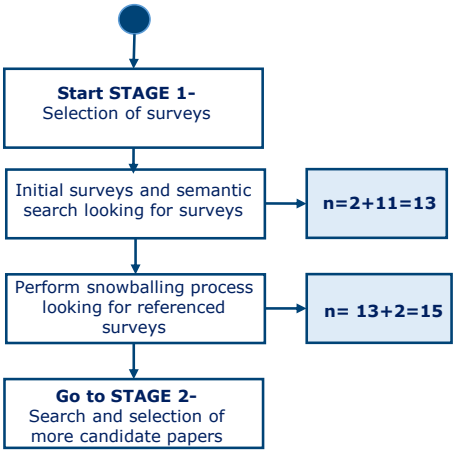

Fig. 2. First review process. Selection of surveys and reviewing papers

by the bibliography reference manager JabRef [11] which allowed the team to manage more efficiently the duplicate references among search sources and stages.

Stage 1- Selection of Surveys and Reviewing Papers

Since we considered published surveys and reviewing papers (from now on we refer to these types of studies as simply *surveys*) on the topic as potentially relevant for the study, we started our search process with the *Stage 1* (see Figure 2) in which we selected surveys published on the topic of provenance. More specifically, first we performed a semantic search (using particularly the semantic searcher *SemanticScholar*) looking for such a kind of literature. These resources would provide us with (1) analysis and comparisons of potentially useful existing research, which could be candidate papers of our systematic review, and (2) relevant conference/workshop proceedings and journals where such an existing research on the topic was published, which we could use as source venues to scan and look for additional research on the topic.

The semantic search was undertaken by Sáenz and Pérez who identified a total number of 13 surveys ([12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], plus the already known surveys ([6], [7]). Taking these surveys as a starting point, we performed a snowball process looking for other surveys, identifying another 2 papers ([23], [24]). Part of the information extracted from each survey is shown in Table 3, where the surveys have been classified in three sections depending on the moment in which they were identified. In this table we particularly present, for each paper, an identifier (*Id*), together with its reference (*Ref.*), a representative name for the authors (*Authors' Rep.Nam.*), and title (*Title*), the venue in which it was published (*Venue*), and the year (*Year publi.*). The remainder data in Table 3 will be explained in the next subsection. These studies, together with our personal knowledge, provided us with the information needed to guide the following stage in our search and selection process.

Stage 2- Search and Selection of More Candidate Papers

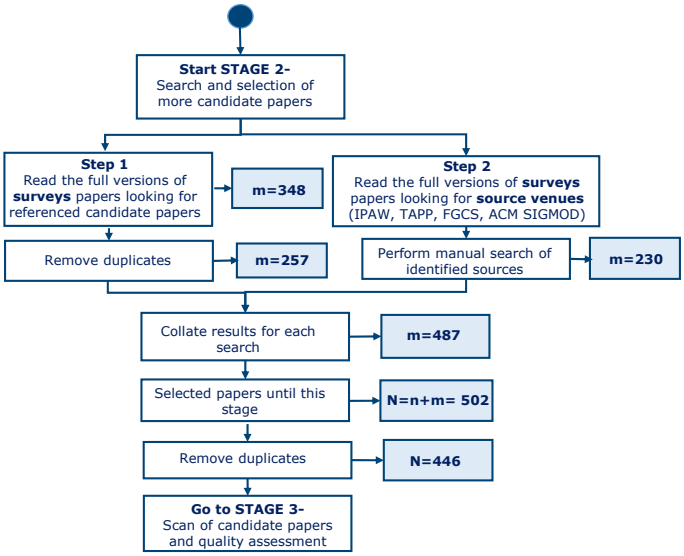

Fig. 3. First review process. Search and selection of more candidate papers

An overview of *Stage 2* is depicted in Figure 3. In this stage we scanned the full versions of the identified surveys looking for two types of information. On one hand, we searched for candidate papers referenced in the surveys, which corresponds to *Step 1* of this stage. On the other hand, we looked for relevant conference/workshop proceedings and journals in which such candidate papers were published. The selected venues would be taken as a starting point to search for new candidate papers, which corresponds to *Step 2* of this stage.

Step 1 This step was undertaken by Sáenz and Pérez which identified any paper referenced in the surveys which addressed any literature regarding provenance (provenance systems, characteristics or features, techniques, etc.). As shown in Table 3, the different surveys covered an interesting range of relevant papers (see column *Num. pap.*) published on overlapping periods from 1986 to 2010 (see column *Range years*), being the stretch of time from 2010 to 2014 only tackled in [*Su1*] [6]. This aspect would imply missing candidate papers published within that period, which we took into account in the following step of *Stage 2*. The results of the searches performed in this step by the two authors were collated and any paper we disagreed about were read and then discussed until finding an agreement (the third author was included to reach a consensus when it was necessary). As a result, we classified a total of 348 works and later we performed a process of removal of duplicates obtaining 257 papers (see Figure 3).

Step 2 Regarding the conference/workshop proceedings and journals, we identified two provenance-specific workshops which held on the topic. These workshops are the biennial International Provenance & Annotation Workshop, and the annual workshop on Theory and Practice of Provenance (TaPP). Additionally, there were other two sources which we

TABLE 3
First review process. Surveys and reviewing papers identified

Id	Ref.	Authors' Rep.Nam.	Title	Venue	Year publi.	Range years	Num. pap.
Su1	[6]	Carata et al.	A Primer On Provenance	Communications of the ACM	2014	2000-2014	14
Su2	[7]	Moreau	The foundations for provenance on the web	Foundations and Trends in Web Science	2010	1986-2009	88
Su3	[12]	Bose and Frew	Lineage Retrieval for Scientific Data Processing. A survey	ACM Computing Surveys (CSUR)	2005	1988-2004	17
Su4	[13]	Buneman et al.	Why and where-A characterization of data provenance	Proc. of the 8th International Conference on Database Theory (ICDT)	2001	1988-2001	2
Su5	[14]	Buneman and Tan	Provenance in Databases	Proc. of the ACM SIGMOD international conference on Management of data (MOD)	2007	1990-2007	15
Su6	[15]	Cheney et al.	Provenance in databases: Why, how, and where	Foundations and Trends in Databases	2009	1981-2009	24
Su7	[16]	Davidson and Freire	Provenance and scientific workflows-challenges and opportunities	Proc. of the ACM SIGMOD MOD	2008	1995-2008	40
Su8	[17]	Davidson et al.	Provenance in Scientific Workflow Systems	IEEE Data Eng. Bull.	2007	1995-2008	14
Su9	[18]	Freire et al.	Provenance for Computational Tasks-A Survey	Computing in Science and Engineering	2008	2005-2008	35
Su10	[19]	Glavic and Dittrich	Data Provenance-A Categorization of Existing Approaches	Proc. of Datenbanksysteme in Business, Technologie und Web (BTW)	2007	1982-2006	28
Su11	[20]	Simmhan et al.	A Survey of data provenance in e-science	ACM SIGMOD Record	2005	1991-2005	27
Su12	[21]	Simmhan et al.	A Survey of Data Provenance Techniques	Technical Report, Computer Science Department, Indiana University	2005	1988-2005	72
Su13	[22]	Tan	Provenance in Databases: Past, Current, and Future	IEEE Data Eng. Bull.	2007	1990-2007	23
Su14	[23]	Buneman and Davidson	Data provenance- the foundation of data quality	Carnegie Mellon University Software Engineering Institute	2010	2004-2010	13
Su15	[24]	Serra et al.	Towards a Taxonomy of Provenance in Scientific Workflow Management Systems	Proc. of the IEEE Congress on Services, Part I (SERVICES)	2009	1997-2009	24
				Total with duplicates			348
				Total without duplicates			**257**

selected because of the high number of papers they published on the topic: the Future Generation Computer Systems (FGCS) journal and several ACM SIGMOD resources (including the SIGMOD Record journal, as well as the ACM SIGMOD International Conference on Management of Data (MOD) and the ACM SIGMOD Symposium on Principles of Database Systems (PODS)). Since the identified surveys tackle relevant papers from 1986 to 2014, but only [*Su1*] [6] considered candidate papers published from 2010 to 2014, we decided to limit the search in these venues from 2010 to 2015, both included.

Regarding the IPAW and TaPP workshops, Sáenz undertook an independent manual search of the different editions, identifying 171 published papers, being 75 in IPAW and 96 in TaPP (see Table 4). Regarding FGCS and the ACM SIGMOD resources, Pérez performed an automatic search choosing Scopus as data source because of its good coverage and reputation. Additionally, given the bounded search considering publications in FGCS and ACM SIGMOD, in order to be as unbiased as possible, we chose the keywords 'provenance' and 'lineage' looking for any paper containing any of these words in the title, abstract or keywords. This automatic search resulted in 59 papers, being 26 published in FGCS and 33 published in ACM SIGMOD resources (see Table 4). The searches in the selected venues identified a total of 230 papers (see Table 4 and Step 2 in Figure 3).

Finally, papers in both the *Step 1* inclusion set and the *Step 2* inclusion set were collated and removed duplicates, resulting in 487 papers. These papers together with the 15 surveys identified in *Stage 1*, reached a total of 502 candidate papers which, after eliminating duplicates, resulted in a total of 446 papers (see Figure 3), giving way to *Stage 3*.

Stage 3- Scan of Candidate Papers and Quality Assessment

In *Stage 3* all the authors went through the studies resulting from *Stage 2* looking for evidence of concrete existing provenance systems, by applying the inclusion and exclusion criteria. In particular, Sáenz addressed papers identified at *Step 1* of *Stage 2*, Pérez was responsible for scanning papers identified at *Step 2* of *Stage 2*, while Rubio was in charge of addressing papers resulted from both steps, so that all papers were scanned by two authors. Facing the wide number of provenance systems existing within the selected studies, we decided to assess the quality of the provenance systems to be chosen for the review, by measuring somehow their impact within the provenance community. More

TABLE 4
Information about the searches in chosen venues

Venue	Acronym	Type	Search period/ Conf. year	Num. papers	Total
Future Generation Computer Systems	FGCS	Journal	from 2010 to 2015	26	26
ACM SIGMOD Record	SIGMOD-Rec	Journal	from 2010 to 2015	4	33
ACM SIGMOD PODS/MOD	SIGMOD-Con	Conference/ Symposium	from 2010 to 2015	29	
Proceedings International Provenance & Annotation Workshop	IPAW	Workshop	2010	39	75
			2012	18	
			2014	18	
Proceedings Theory and Practice of Provenance	TAPP	Workshop	2010	11	96
			2011	29	
			2012	16	
			2013	12	
			2014	14	
			2015	14	
		Total without duplicates			**230**

specifically, we considered that a study cites a provenance system either for its importance or interest or for being somehow related to the proposal presented in such a study. Based on this premise, we decided to scan each study from these identified until this moment, registering the provenance system presented in such a study (if applicable) as well as any other provenance system cited on it. As a result of this process, we identified a total of 170 different provenance systems, being 23 by far the most referenced for their repercussion and consolidation within the provenance context (in Appendix A, the reader can find the list of the provenance systems identified in the papers resulted from *Stage 2*, together with their number of appearances). Taking this into account, we decided to apply the selection filter of the papers identified from *Stage 2*, looking for (1) those which specially tackled the chosen systems, or (2) perform any type of quality analysis of the systems (such as surveys or reviewing papers).

In the particular case of duplicate publications, in the case in which a work had conference/workshop and extended journal versions, we manually checked such versions and we only included in the final set of primary papers, journal papers or articles with more details of the study. Though it was not in our selection criteria in protocol, during quality assessment and data extraction phase we came across some extremely low quality and plagiarized papers (in both cases they were irrelevant). We decided to exclude them. This process was performed by all the authors which carried out the process independently and, after finishing, analyzed in consensus meetings possible disagreements resolving them by discussing the matter until agreement was reached. As a result, a total of 64 papers were chosen (see Figure 4).

Starting from these 64 papers, we scanned their reference lists in order to identify further papers; in particular, we undertook a process of snowballing, paying special attention to "Related Work" sections, and complete texts of surveys or reviewing papers that analyze provenance systems. This step identified 9 new papers, achieving a total of 73 papers (see Figure 4). Additionally, as another way of quality assessment, we contacted by email the corresponding authors of the chosen systems (23 distinct research groups) in order to explore whether they were aware of relevant papers not

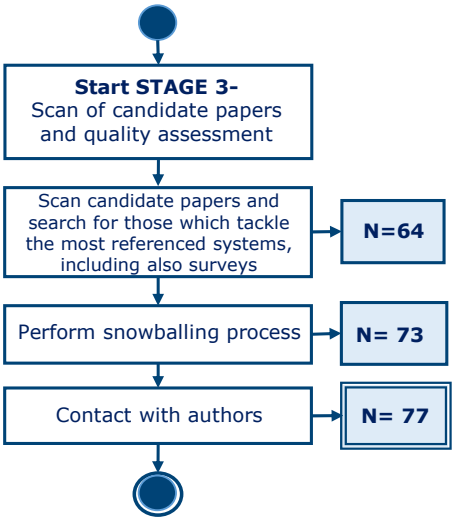

Fig. 4. First review process. Scan of candidate papers and quality assessment

included in our list. The authors' replies resulted in adding 4 articles, so that the final list of papers included 77 articles (see Figure 4). Two of the four suggested papers were not initially included as part of our set of selected papers since they presented a recent extension of a provenance system already included in our study, but such an extension had had little publication impact to be included in our list of most relevant provenance systems. We decided to include these papers in the review, considering both systems. The other two papers were not detected in the earlier stages of the search because they were published on-line, but not in a regular issue (they particularly correspond to a technical report and a PhD thesis).

As a result, as it can be inferred from the selected papers, they can be classified into two different groups attending to (1) those papers which refer to surveys (15 papers), and (2) those papers that tackle any of the concrete provenance systems classified for our review (62 papers).

TABLE 5
First review process. Selected studies regarding the concrete systems

Id	System name	Authors' Rep.Nam.	Ref. *Stage 3-1*	Ref. *Stage 3-2*	Ref. *Stage 3-3*
S1	VisTrails	J. Freire et al.	[25], [26]		
S2	myGrid/Taverna	K. Wolstencroft et al.	[27]		
S3	Kepler-provenance	I. Altintas et al.	[28]		
S4	PASS	D. A. Holland et al.	[29], [30]	[31]	
S5	Trio	J. Widom et al.	[32], [33], [34], [35], [36]		[37]
S6	Karma	L.Y. Simmhan et al.	[38], [39]		
S7	Chimera	I. Foster et al.	[40], [41], [42]		
S8	ZOOM	O. Biton et al.	[43], [44], [45], [46]		
S9	Cui 2000	Y. Cui et al.	[47], [48], [49]	[50]	
S10	Swift	I. Foster et al.	[51], [52]		
S11	PASOA/PreServ	P. Groth et al.	[53]		
S12	Tioga	M. Stonebraker et al.	[54], [55], [56]		
S13	Wings-Pegasus	E. Deelman et al.	[57], [58]	[59], [60], [61]	
S14	Buneman	P. Buneman et al.	[62]		
S15	SPADE	A. Gehani et al.	[63]	[64]	
S16	ORCHESTRA	Z.G. Ives et al.	[65], [66], [67]		
S17	Perm-GProM	B. Glavic et al.	[68], [69], [70], [71], [72]	[73], [74], [75]	[76], [77]
S18	ES3	J. Frew et al.	[78], [79]		
S19	COMAD	S. Bowers et al.	[80]		
S20	DBNotes	L. Chiticariu et al.	[81], [82]		
S21	LipStick	Y. Amsterdamer et al.	[83]		
S22	REDUX	R. Barga et al.	[84]		
S23	BURRITO	P.J. Guo et al.	[85]		[86]

TABLE 6
General data items extracted from each paper

Data items	Description
Identifier	Unique identifier for the paper (same as the bibliographic reference number).
Bibliographic	Author(s), year, title, source, Web-site (if possible).
Type of article	Journal/ Conference/ Workshop/ PhD thesis/ Books/ Technical report/ White paper
Related papers by authors	References of related works by the same authors or the same research group.

2.4 Data Extraction

In this section, we explain the strategy followed to extract data from the papers selected during the first review process. Firstly, we identified standard information to be obtained from each selected paper, such as a unique identifier of each paper, bibliographical aspects, the type of article and papers related by authors. This information is shown in Table 6. Secondly, given the particularity of our research questions, we performed a two-phase process of data extraction depending on the research question to be tackled.

On one hand, research questions *RQ1* and *RQ2* are closely related among them since while *RQ1* aims to identify the different aspects in which existing provenance systems could be categorized, *RQ2* is devoted to characterizing the techniques or solutions each provenance proposal provides to address such aspects. Taking this into account, in order to give an answer to questions *RQ1* and *RQ2*, we took as source all the selected papers, paying special attention to the surveys (see Table 3) since these type of works explicitly describe and/or categorize provenance systems. Additionally, we established a data extraction form aimed to identify from the selected papers:

- any classification of aspects authors considered to compare the systems and proposals tackled on it (related to *RQ1*).
- the techniques considered by the systems to address such different aspects (related to *RQ2*).

In particular, the data collection form was set up in a word-processing document. The data extraction process in this case was finalized after the three authors trialed the data extraction on the overall papers.

On the other hand, as explained previously, providing an answer to question *RQ3* requires having previously answered *RQ1* and *RQ2*, that is, having established the different aspects or categories in base on which the selected provenance systems must be analyzed. For this reason, the data extraction form needed to answer *RQ3* was established after obtaining the answers to questions *RQ1* and *RQ2*. In particular, such a form mainly defined an extraction item per each aspect identified previously. In this case, the data collection form was set up in a spreadsheet and, based on this sheet, each author undertook an independent collection process from all the selected papers (both those papers which tackle the specific provenance systems and the surveys).

3 The Second Review Process

As advanced previously, the first review process was renewed by performing a second review process which covered research studies published from January 2016 to July 2017.

Starting from the protocol defined in the first review process, in this process we performed the *conducting* and the *reporting* stages. Next, we describe the main phases we performed in this second review process regarding the *conducting* stage. We note that the methodology followed to perform such a stage was the same than the one taken during the first review process regarding what authors performed what steps, how we resolved the possible disagreements, etc.

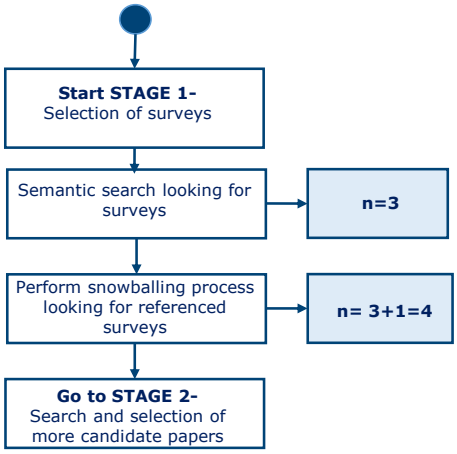

Fig. 5. Second review process. Selection of surveys and reviewing papers

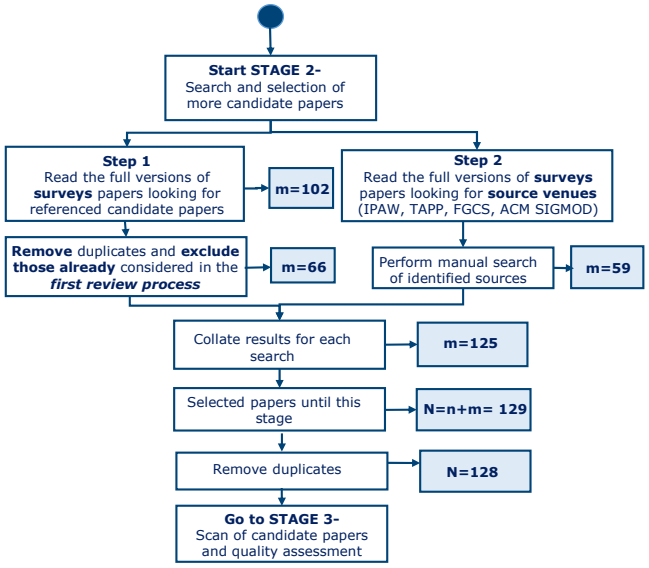

Fig. 6. Second review process. Search and selection of more candidate papers

As in the previous review process, the results obtained from the data synthesis of the studies selected in this second review process are presented in the original paper [4].

3.0.1 Search and Selection Strategy

Given the experience gained in the first review process, in this process we again combined an automatic search with a manual search. More specifically, the search and selection process was performed in three stages, as outlined in Figures 5-7 respectively, and which are described in more detail in the following subsections.

Stage 1- Selection of Surveys and Reviewing Papers

As depicted in Figure 5, in this stage we firstly performed a semantic search on *SemanticScholar* looking for surveys and reviewing papers published from January 2016 to July 2017. The search identified 3 new surveys ([87], [88], [89]). Taking these surveys as a starting point, we performed a snowball process looking for other surveys, identifying another paper ([90]). Part of the information extracted from each survey is shown in Table 7. We want to note that, since the survey [Su17] was not available online, we need to contact by email its author in order to get the paper.

Stage 2- Search and Selection of More Candidate Papers

In this stage (see Figure 6) we again performed two steps: (*Step 1*) searching for candidate papers referenced in the 4 surveys identified previously, and (*Step 2*) looking for relevant conference/workshop proceedings and journals in which such candidate papers were published.

Step 1 In columns *Num. pap.* and *Range years* of Table 7, we show the number of relevant papers referenced in these 4 surveys, together with their range of publication years, respectively. Overall, we classified a total of 102 works. Later, we performed a process of removal of duplicates, resulting in a list of 97 papers. At this point, we needed to perform an additional task. Among the papers selected until this moment, there were papers that, for their interest, could have been already considered in the first review process. For this reason, we checked the 97 papers looking for those which matched up with any of the candidate papers considered in the first review process (446 papers), removing them from our list of papers. As a result, we obtained a list of 66 papers (see Table 7 and Step 1 in Figure 6).

Step 2 Taking into account the venues of publication of the relevant papers referenced in the selected surveys, we decided to choose the same source venues as in the first review process (FGCS, SIGMOD-Rec, SIGMOD-Con, IPAW, TAPP). Since this second review process aimed at covering research studies published from January 2016 to July 2017, we limited the search in these venues to such a period.
In this stage, we first performed a manual search in the IPAW and TaPP workshops. Since IPAW is biennial and Tapp is annual, we searched in the editions IPAW'16, and TaPP'16 and TaPP'17 (see Table 8). More specifically, we identified 36 published papers, being 14 in IPAW and 22 in TaPP (see Table 8). Additionally, we performed an automated search in FGCS and the ACM SIGMOD resources using the keywords 'provenance' and 'lineage' and looking for any paper containing any of these words in the title, abstract or keywords. The search resulted in 23 papers, being 14 published in FGCS and 9 published in ACM SIGMOD resources (see Table 8). We note that we did not get any paper published in SIGMOD-Rec verifying the search's constraints. The searches in the selected venues identified a total of 59 papers (see Table 8 and Step 2 in Figure 6). Given the recent celebration of the conferences and workshops considered in this search, we again needed to contact by email the authors of two papers selected in this step, since they were not available online.

After collating the papers obtained from *Step 1* and *Step 2*, we obtained 125 papers (see Figure 6). These papers

TABLE 7
Second review process. Identified surveys and reviewing papers

Id	Ref.	Authors' Rep.Nam.	Title	Venue	Year publi.	Range years	Num. pap.
Su16	[87]	Cuzzocrea	Big Data Provenance: State-Of-The-Art Analysis and Emerging Research Challenges	Proceedings of the Workshops of the (EDBT/ICDT'16	2016	2006-2015	18
Su17	[88]	Dogan	A Survey of Provenance in Wireless Sensor Networks	Ad Hoc & Sensor Wireless Networks	2016	2000-2014	47
Su18	[89]	Wang et al.	Provenance for Wireless Sensor Networks: A Survey	Data Science and Engineering	2016	2006-2016	12
Su19	[90]	Tan et al.	Security and Data Accountability in Distributed Systems: A Provenance Survey	Proceedings of IEEE HPCC13	2013	2002-2011	24
				Total with duplicates			102
				Total without papers already considered in the first review process and without duplicates			**66**

TABLE 8
Second review process. Information about the searches in chosen venues

Venue	Acronym	Type	Search period/ Conf. year	Num. papers	Total
Future Generation Computer Systems	FGCS	Journal	from 2016 to July 2017	14	14
ACM SIGMOD PODS/MOD	SIGMOD-Con	Conference/ Symposium	from 2016 to July 2017	9	9
Proceedings International Provenance & Annotation Workshop	IPAW	Workshop	2016	14	14
Proceedings Theory and Practice of Provenance	TAPP	Workshop	2016	11	
			2017	11	22
		Total without duplicates			**59**

together with the 4 surveys identified in *Stage 1*, reached a total of 129 candidate papers. Finally, after eliminating duplicates, we obtain a total of 128 papers (see Figure 6), giving way to *Stage 3*.

Stage 3- Scan of Candidate Papers and Quality Assessment

In *Stage 3* we first needed to know whether there were new provenance systems of interest to be analyzed. For this task, we decided to scan each study from these identified until this moment in this review process (128 papers), registering the provenance system presented in such a study (if applicable) as well as any other provenance system cited on it. We included this information to the ranking of provenance systems identified in the first review process (presented in Appendix A). As a result of this task, we extended the previous list of provenance systems from 170 to 251 (in Tables 14-16 of Appendix B, the reader can find the list of the provenance systems we identified in this second review process together with their number of appearances). As we show in the list of provenance systems of Appendix B, in this second review process the systems PLUS and RAMP climbed positions in the list, being above Redux and Burrito (see Table 14), analyzed in the first review process. For this reason, we decided to renew the systematic review including also these two new provenance systems, analyzing a total of 25 systems (see grey cells in Table 14 of Appendix B).

Second, having chosen the provenance systems to be analyzed, we needed to look for (i) those papers which specially tackled such systems, or (ii) perform any type of quality analysis of the systems (such as surveys or reviewing papers). At this point, we would like to note that the set of papers in which we needed to apply the selection filters was composed by two groups:

- the papers identified from *Stage 2* of this second review process, where we looked for information regarding any of the selected 25 systems (we refer to the search performed in these papers as *Step 1* of the *Stage 3*).
- the papers identified from *Stage 2* of the first review process, where we looked for information regarding the two new systems exclusively. We refer to the search performed in these papers as *Step 2* of the *Stage 3*. We perform this second step because in the first review process PLUS and RAMP were not included among our chosen systems, however the papers identified from *Stage 2* of such review process were potential candidate papers to contain interesting information of such systems.

In Figure 7, we show the two steps. From the *Step 1* of this stage we identified 4 papers and, after performing a process of snowballing, we identified another 4 papers (8 papers in total). We want to note that in the snowballing process we paid special attention to "Related Work" sections, and complete texts of surveys or reviewing papers that analyze provenance systems. From the *Step 2* of this stage we identified 17 papers and, after undertaking a snowballing process, we selected another 7 papers (24 papers in total).

Finally, we collated the results obtained from the two steps, removed duplicates and papers already considered in the first review process, achieving a total of 28 papers (see Figure 7).

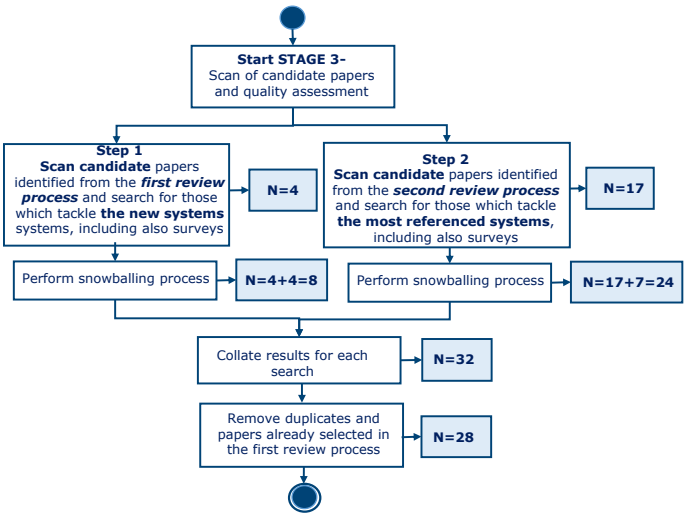

Fig. 7. Second review process. Scan of candidate papers and quality assessment

As a result, we can again classify the papers selected from this second review process attending to (1) those papers which refer to surveys (4 papers), and (2) those papers that tackle any of the concrete provenance systems classified for our review (24 papers). In particular, the selected studies regarding the concrete systems are shown in Table 9.

Regarding the *Data Extraction* task, we note that the strategy followed in this second review process was the same than the chosen in the first review process.

Finally, the overall works selected in both review processes, regarding the chosen surveys and the concrete systems, are presented in Tables 10 and 11. Additionally, in Appendix C we also show a brief description of the surveyed systems.

4 Discussion

In this section, we first discuss the limitations of this review, and go on to analyze the threats to validity arising from the procedures we followed to perform the systematic review.

4.1 Study Limitations

Despite the fact that we try our best to adhere to the guidelines [1], [2], [3], we can still identify some limitations to our study. With respect to the search process followed in both review processes, given the huge number of results obtained from our automatic search described in Subsection 2.3.1, and checking practices of other researchers looking at research trends [3], [10], we decided to perform an alternative search strategy that basically combined automatic and manual searches. Although with such a strategy we have checked references of a wide number of articles as well as many venues, our search of relevant studies may have not been thorough. In the review, we have restricted ourselves to English–language studies and limited the time span of the studies only to those published up to and including July 2017. Overall, though, we do not expect to have missed a large number of important studies.

As regards data extraction, we should also mention that some aspects of the literature, related to the application of our taxonomy to the surveyed provenance systems, have been rooted on our own interpretation of the papers (especially in these situations in which the strategy was not clearly described). For this reason, it is possible that other researchers might arrive at different conclusions.

4.2 Threads to Validity

A systematic literature review such as this one has several evident threats to its validity. First, whether we have chosen enough search resources. On this count, the ample list of different papers indicates that the width of the search is sufficient.

Second, another possible threat to validity is bias in applying quality assessment and data extraction. In order to minimize this threat insofar as possible, we explicitly defined the inclusion and exclusion criteria, which we believe was detailed enough to provide an assessment of how we selected the chosen papers for analysis.

Finally, another important threat to validity is reliability, which focuses on whether the data are extracted and the analysis is performed in such a way that it can be repeated by other researchers to obtain the same results. In relation to this, we have explained in detail the search process of both review processes and procedures applied during the overall review so that it may be replicated by others, with the exception of, as described in Subsection 4.1, those aspects considered in the application of our taxonomy to the different systems, where in some situations we have based on our own interpretation of the papers.

Appendix A
Identified Provenance Systems in the First Review Process

In Tables 12 and 13 we show the set of provenance systems distinguished from the papers identified from *Stage 2* of the search and selection process of the first review process. More specifically, per each system we show its name, including its citation when no name is given, together with the number of appearances. It is important to note that the systems identified with numbers from 1 to 23 (see grey cells in Table 12) correspond to the provenance systems we analyzed in the first review process of the systematic review.

Appendix B
Identified Provenance Systems in the Second Review Process

Tables 14 and 15 show the set of provenance systems identified after performing both the first and second review processes. It is important to note that the systems identified with numbers 22 and 23 correspond to the new provenance systems we included in the second review process.

Appendix C
A Brief Introduction of Surveyed Systems

Table 17 shows a brief description of the surveyed systems.

TABLE 9
Second review process. Selected studies regarding the concrete systems

Id	System name	Authors' Rep.Nam.	Ref. *Stage 3-1*	Ref. *Stage 3-2*
S1	VisTrails	J. Freire et al.	[91]	[92], [93]
S2	myGrid/Taverna	K. Wolstencroft et al.	[94], [95]	[96]
S3	Kepler-provenance	I. Altintas et al.	[97]	
S4	PASS	D. A. Holland et al.		
S5	Trio	J. Widom et al.		
S6	Karma	L.Y. Simmhan et al.	[98]	[99]
S7	Chimera	I. Foster et al.		
S8	ZOOM	O. Biton et al.		
S9	Cui 2000	Y. Cui et al.		
S10	Swift	I. Foster et al.	[100]	[101]
S11	PASOA/PreServ	P. Groth et al.		
S12	Tioga	M. Stonebraker et al.		
S13	Wings-Pegasus	E. Deelman et al.	[102]	[103]
S14	Buneman	P. Buneman et al.		
S15	SPADE	A. Gehani et al.	[104], [105], [106]	
S16	ORCHESTRA	Z.G. Ives et al.	[107]	[108]
S17	Perm-GProM	B. Glavic et al.		
S18	ES3	J. Frew et al.		
S19	COMAD	S. Bowers et al.		
S20	DBNotes	L. Chiticariu et al.		
S21	LipStick	Y. Amsterdamer et al.		
S22	REDUX	R. Barga et al.		
S23	BURRITO	P.J. Guo et al.		
S24	PLUS	B. Blaustein et al.	[109]	[110], [111], [112]
S25	RAMP	R. Ikeda	[113], [114]	

REFERENCES

[1] B. Kitchenham, "Procedures for Performing Systematic Reviews," Keele University, Technical Report TR/SE-0401 (2004), at: http://www.inf.ufsc.br/ aldo.vw/kitchenham.pdf. Last visited on August 2017.

[2] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," no. EBSE 2007–01, at: http://pages.cpsc.ucalgary.ca/∼sillito/cpsc-601.23/readings/kitchenham-2007.pdf. Visited on August 2017.

[3] B. Kitchenham, O. P. Brereton, D. Budgen *et al.*, "Systematic Literature Reviews in Software Engineering - A Systematic Literature Review," *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, Jan. 2009.

[4] B. Pérez, J. Rubio, and C. Sáenz-Adán, "A systematic review of provenance systems," *Submitted to Knowledge and Information Systems*, 2017.

[5] H. Zhang, M. A. Babar, and P. Tell, "Identifying Relevant Studies in Software Engineering," *Inf. Softw. Technol.*, vol. 53, no. 6, pp. 625–637, Jun. 2011.

[6] L. Carata, S. Akoush, N. Balakrishnan *et al.*, "A Primer on Provenance," *Communications of the ACM*, vol. 12, no. 3, pp. 10:10–10:23, Mar. 2014.

[7] L. Moreau, "The Foundations for Provenance on the Web," *Foundations and Trends in Web Science*, vol. 2, no. 2–3, pp. 99–241, 2010.

[8] M. Petticrew and H. Roberts, *Systematic Reviews in the Social Sciences: A Practical Guide*. Blackwell Publishing, 2005.

[9] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, "Lessons from Applying the Systematic Literature Review Process Within the Software Engineering Domain," *J. Syst. Softw.*, vol. 80, no. 4, pp. 571–583, 2007.

[10] D. I. K. Sjoberg, J. E. Hannay, O. Hansen, V. By Kampenes, A. Karahasanovic, N.-K. Liborg, and A. C. Rekdal, "A Survey of Controlled Experiments in Software Engineering," *IEEE Trans. Softw. Eng.*, vol. 31, no. 9, pp. 733–753, Sep. 2005.

[11] JabRef, available at: http://www.jabref.org/. Last visited on August 2017.

[12] R. Bose and J. Frew, "Lineage retrieval for scientific data processing: a survey," *ACM Comput. Surv.*, vol. 37, no. 1, pp. 1–28, 2005.

[13] P. Buneman, S. Khanna, and W. C. Tan, "Why and Where: A Characterization of Data Provenance," in *Proc. of the 8th Intl. Conf. on Database Theory (ICDT)*, 2001, pp. 316–330.

[14] P. Buneman and W. C. Tan, "Provenance in Databases," in *Proc. of MOD'07*. ACM, 2007, pp. 1171–1173.

[15] J. Cheney, L. Chiticariu, and W. C. Tan, "Provenance in Databases: Why, How, and Where," *Foundations and Trends in Databases*, vol. 1, no. 4, pp. 379–474, Apr. 2009.

[16] S. B. Davidson and J. Freire, "Provenance and scientific workflows: challenges and opportunities," in *Proc. of MOD'08*, 2008, pp. 1345–1350.

[17] S. B. Davidson, S. Cohen-Boulakia, A. Eyal *et al.*, "Provenance in Scientific Workflow Systems," *IEEE Data Eng. Bull.*, vol. 30, no. 4, pp. 44–50, 2007.

[18] J. Freire, D. Koop, E. Santos, and C. T. Silva, "Provenance for Computational Tasks: A Survey," *Comput. Sci. Eng.*, vol. 10, no. 3, pp. 11–21, May 2008.

[19] B. Glavic and K. R. Dittrich, "Data Provenance: A Categorization of Existing Approaches," in *Proc. of Datenbanksysteme in Business, Technologie und Web (BTW'07)*, 2007, pp. 227–241.

[20] Y. L. Simmhan, B. Plale, and D. Gannon, "A Survey of Data Provenance in e-Science," *SIGMOD Rec.*, vol. 34, no. 3, pp. 31–36, 2005.

[21] ——, "A Survey of Data Provenance Techniques," Tech. Rep. 612, extended version of SIGMOD Record 2005. Available at: http://www.cs.indiana.edu/pub/techreports/TR618.pdf. Last visited on August 2017.

[22] W. C. Tan, "Provenance in Databases: Past, Current, and Future," *IEEE Data Eng. Bull.*, vol. 30, no. 4, pp. 3–12, 2007.

[23] P. Buneman and S. B. Davidson , "Data provenance – the foundation of data quality," available at: www.sei.cmu.edu/measurement/research/upload/Davidson.pdf. Last visited on August 2017.

[24] S. M. S. da Cruz, M. L. M. Campos, and M. Mattoso, "Towards a Taxonomy of Provenance in Scientific Workflow Management Systems," in *Proc. of the IEEE Congress on Services, Part I, SERVICES I*, 2009, pp. 259–266.

[25] L. Bavoil, S. P. Callahan, P. J. Crossno *et al.*, "Vistrails: Enabling interactive multiple-view visualizations," in *Proc. of the IEEE Visualization (VIS'05)*. IEEE, 2005, pp. 135–142.

[26] J. Freire, C. T. Silva, S. P. Callahan *et al.*, "Managing rapidly-evolving scientific workflows," in *Proc. of IPAW'06*. Springer, 2006, pp. 10–18.

[27] K. Wolstencroft, R. Haines, D. Fellows *et al.*, "The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud," *Nucleic acids research*, pp. 557–561, 2013.

[28] I. Altintas, O. Barney, and E. Jaeger-Frank, "Provenance Collection Support in the Kepler Scientific Workflow System," in *Proc. of IPAW'06*, 2006, pp. 118–132.

[29] K. K. Muniswamy-Reddy, D. A. Holland, U. Braun, and M. I. Seltzer, "Provenance-Aware Storage Systems," in *USENIX Annual Technical Conference, General Track*, 2006, pp. 43–56.

TABLE 10
Overall selected surveys

Id	Survey's name	Authors' Rep.Nam.	References
Su1	A Primer On Provenance	Carata et al.	[6]
Su2	The foundations for provenance on the web	Moreau	[7]
Su3	Lineage Retrieval for Scientific Data Processing. A survey	Bose and Frew	[12]
Su4	Why and where-A characterization of data provenance	Buneman et al.	[13]
Su5	Provenance in Databases	Buneman and Tan	[14]
Su6	Provenance in databases: Why, how, and where	Cheney et al.	[15]
Su7	Provenance and scientific workflows-challenges and opportunities	Davidson and Freire	[16]
Su8	Provenance in Scientific Workflow Systems	Davidson et al.	[17]
Su9	Provenance for Computational Tasks-A Survey	Freire et al.	[18]
Su10	Data Provenance-A Categorization of Existing Approaches	Glavic and Dittrich	[19]
Su11	A Survey of data provenance in e-science	Simmhan et al.	[20]
Su12	A Survey of Data Provenance Techniques	Simmhan et al.	[21]
Su13	Provenance in Databases: Past, Current, and Future	Tan	[22]
Su14	Data provenance ? the foundation of data quality	Buneman and Davidson	[23]
Su15	Towards a Taxonomy of Provenance in Scientific Workflow Management Systems	Serra et al.	[24]
Su16	Big Data Provenance: State-Of-The-Art Analysis and Emerging Research Challenges	Cuzzocrea	[87]
Su17	A Survey of Provenance in Wireless Sensor Networks	Dogan	[88]
Su18	Provenance for Wireless Sensor Networks: A Survey	Wang et al.	[89]
Su19	Security and Data Accountability in Distributed Systems: A Provenance Survey	Tan et al.	[90]

TABLE 11
Selected studies regarding the concrete systems

Id	System name	Authors' Rep.Nam.	References
S1	VisTrails	J. Freire et al.	[25], [26], [91], [92], [93]
S2	myGrid/Taverna	K. Wolstencroft et al.	[27], [94], [95], [96]
S3	Kepler-provenance	I. Altintas et al.	[28], [97]
S4	PASS	D. A. Holland et al.	[29], [30], [31]
S5	Trio	J. Widom et al.	[32], [33], [34], [35], [36] [37]
S6	Karma	L.Y. Simmhan et al.	[38], [39] [98], [99]
S7	Chimera	I. Foster et al.	[40], [41], [42]
S8	ZOOM	O. Biton et al.	[43], [44], [45], [46]
S9	Cui 2000	Y. Cui et al.	[47], [48], [49], [50]
S10	Swift	I. Foster et al.	[51], [52] [100], [101]
S11	PASOA/PreServ	P. Groth et al.	[53]
S12	Tioga	M. Stonebraker et al.	[54], [55], [56]
S13	Wings-Pegasus	E. Deelman et al.	[57], [58], [59], [60], [61] [102], [103]
S14	Buneman	P. Buneman et al.	[62]
S15	SPADE	A. Gehani et al.	[63], [64] [104], [105], [106]
S16	ORCHESTRA	Z.G. Ives et al.	[65], [66], [67], [107], [108]
S17	Perm-GProM	B. Glavic et al.	[68], [69], [70], [73], [74] [71], [72], [75], [76], [77]
S18	ES3	J. Frew et al.	[78], [79]
S19	COMAD	S. Bowers et al.	[80]
S20	DBNotes	L. Chiticariu et al.	[81], [82]
S21	LipStick	Y. Amsterdamer et al.	[83]
S22	REDUX	R. Barga et al.	[84]
S23	BURRITO	P.J. Guo et al.	[85] [86]
S24	PLUS	B. Blaustein et al.	[109], [110], [111], [112]
S25	RAMP	R. Ikeda	[113], [114]

[30] K. K. Muniswamy-Reddy, U. Braun, D. A. Holland *et al.*, "Layering in Provenance Systems," in *USENIX Annual Technical Conference*, 2009.

[31] D. A. Holland, M. I. Seltzer, U. Braun, and K. K. Muniswamy-Reddy, "PASSing the provenance challenge," *Concurr Comput*, vol. 20, no. 5, pp. 531–540, 2008.

[32] J. Widom, "Trio: A System for Integrated Management of Data, Accuracy, and Lineage," Stanford InfoLab, Technical Report 2004-40, available at: http://ilpubs.stanford.edu:8090/658/. Last visited on August 2017.

[33] P. Agrawal, O. Benjelloun, A. D. Sarma *et al.*, "Trio: A System for Data, Uncertainty, and Lineage," in *Proc. of VLDB'06*, September 2006, pp. 1151–1154.

[34] O. Benjelloun, A. D. Sarma, C. Hayworth, and J. Widom, "An Introduction to ULDBs and the Trio System," Stanford InfoLab, Technical Report 2006-7, avail. at: http://ilpubs.stanford.edu:8090/793/. Last visited on August 2017.

[35] M. Mutsuzaki, M. Theobald, and et al, "Trio-One: Layering Uncertainty and Lineage on a Conventional DBMS," in *Proc. of CIDR'07*, 2007, pp. 269–274.

[36] J. Widom, "Trio: A System for Data, Uncertainty, and Lineage," in *Managing and Mining Uncertain Data*. Springer, 2008.

[37] P. Agrawal, R. Ikeda, H. Park, and J. Widom, "Trio-ER: The Trio System as a Workbench for Entity-Resolution," Stanford University, Technical Report (March 2009), avail. at: http://ilpubs.stanford.edu:8090/912/. Last visited on August 2017.

[38] Y. L. Simmhan, B. Plale, D. Gannon, and S. Marru, "Performance evaluation of the karma provenance framework for scientific workflows," in *Proc. of IPAW'06*, 2006, pp. 222–236.

[39] Y. L. Simmhan, B. Plale, and D. Gannon, "A framework for collecting provenance in data-centric scientific workflows," in *Proc. of the International Conference on Web Services (ICWS'06)*. IEEE, 2006, pp. 427–436.

[40] I. Foster, J. Vöckler, M. Wilde, and Y. Zhao, "Chimera: A virtual data system for representing, querying, and automating data derivation," in *Proc. of SSDBM'02*. IEEE, 2002, pp. 37–46.

[41] B. Clifford, I. Foster, J.-S. Voeckler *et al.*, "Tracking provenance in a virtual data grid," *Concurr Compute*, vol. 20, no. 5, pp. 565–575, 2008.

[42] Y. Zhao, M. Wilde, and I. Foster, "Applying the virtual data provenance model," in *Proc. of IPAW'06*, ser. LNCS, vol. 4145. Springer, 2006, pp. 148–161.

[43] O. Biton, S. Cohen-Boulakia, and S. B. Davidson, "Zoom*UserViews: Querying relevant provenance in workflow systems," in *Proc. of VLDB'07*. VLDB Endowment, 2007, pp. 1366–1369.

[44] S. Cohen-Boulakia, O. Biton, S. Cohen, and S. Davidson, "Addressing the Provenance Challenge using ZOOM," *Concurr Comput*, vol. 20, no. 5, pp. 497–506, 2008.

[45] O. Biton, S. Cohen-Boulakia, S. B. Davidson, and C. S. Hara, "Querying and managing provenance through user views in scientific workflows," in *Proc. of the IEEE 24th International Conference on Data Engineering (ICDE'08)*. IEEE, 2008, pp. 1072–1081.

[46] J. Cheney and R. Perera, "An analytical survey of provenance sanitization," in *Proc. of IPAW'14*, 2014, pp. 113–126.

[47] Y. Cui and J. Widom, "Lineage tracing in data warehouses," in *Proc. of the 16th International Conference on Data Engineering*. IEEE, 2000, pp. 367–378.

[48] Y. Cui, J. Widom, and J. L. Wiener, "Tracing the lineage of view data in a warehousing environment," *ACM Transactions on Database Systems (TODS)*, vol. 25, no. 2, pp. 179–227, 2000.

[49] Y. Cui and J. Widom, "Practical lineage tracing in data warehouses," in *Proc. of the 16th International Conference on Data Engineering (ICDE'00)*. IEEE, 2000, pp. 367–378.

[50] J. Wiener, H. Gupta, W. Labio *et al.*, "A system prototype for warehouse view maintenance," in *Proc. of MOD'95*, 1995, pp. 26–33.

TABLE 12
Provenance Systems identified in the first review process and their appearances(I)

Ref.	System	Num.	Ref.	System	Num.	Ref.	System	Num.	Ref.	System	Num.
1	VisTrails	70	46	pPOD	5	91	IncPy	2	136	PMAF	1
2	myGrid/ Taverna	70	47	CDE	5	92	Gibson et al. [115]	2	137	Sig.ma	1
3	kepler	62	48	Chiron	5	93	PAPEL	2	138	GExpLine	1
4	PASS	51	49	noWorkflow	5	94	Vismashup	2	139	GExpLine	1
5	Trio	22	50	StarFlow	5	95	Story Book	2	140	Matrioshka	1
6	Karma	22	51	ProvBench	4	96	VIEW	2	141	Windows Workflow Foundation	1
7	Chimera	18	52	IWBrowser	4	97	ExSpan	2	142	SAF	1
8	ZOOM	18	53	Geo-Opera	4	98	Gaea	2	143	Tabulator	1
9	Cui 2000	18	54	Marathe	4	99	Geodise	2	144	POMELO	1
10	swift	17	55	S-Language	4	100	Geolineus	2	145	IWBase	1
11	PASOA/PreServ	17	56	TDB	4	101	GOOSE	2	146	Prefuse	1
12	Tioga	17	57	TREC	4	102	Tupelo	2	147	GridSpace	1
13	Wings-Pegasus	16	58	Provenance Explorer	4	103	ViNE	2	148	Vlewfbus	1
14	*Buneman* [62]	16	59	Artemis	4	104	W7	2	149	CAMERA	1
15	SPADE	12	60	LabelFlow	4	105	SELinks	2	150	Condor	1
16	ORCHESTRA	12	61	SCiCumulus	3	106	NetLogo	2	151	eScience Central	1
17	Perm	10	62	Bochner et al. 2008	3	107	Bonsai	2	152	Subdue	1
18	ES3	10	63	IPython	3	108	ReTrace	2	153	SAGA	1
19	COMAD	10	64	OPUS	3	109	Antfarm	2	154	Gspan and close-graph	1
20	DBNotes	9	65	Sahoo et al. [116]	3	110	PubFlow	2	155	Talos	1
21	Lipstick	9	66	ARK 2008	3	111	Caps	1	156	Mars	1
22	REDUX	9	67	DEEP	3	112	Kieker	1	157	CARMEN	1
23	BURRITO	9	68	DataTracker	3	113	Tracefs	1	158	NetTrails	1
24	PLUS	8	69	Knitr	3	114	Sourcetrac	1	159	Amazonia	1
25	probe-it	8	70	Oozie	3	115	IADB	1	160	BDBMS	1

TABLE 13
Provenance Systems identified in the first review process and their appearances (II)

Ref.	System	Num.	Ref.	System	Num.	Ref.	System	Num.	Ref.	System	Num.
26	ParaView	8	71	WDO-It!	3	116	Utopia	1	161	CCDB	1
27	SAM	8	72	ourSpaces	3	117	Dexy	1	162	CRISTAL	1
28	Web Provenance	8	73	RDFProv	3	118	Myers et al. [117]	1	163	ESP2Net	1
29	Provmanager	8	74	RunMyCode.org	3	119	SHARE	1	164	gryPhyN	1
30	PANDA	7	75	LogicBlox	3	120	A-R-E	1	165	LabBase	1
31	Janus	7	76	SystemT	3	121	Collage	1	166	MCA	1
32	SOLE	7	77	ZOO	3	122	Minebench	1	167	MMS	1
33	Git2prov	7	78	DataHub	3	123	Prizms	1	168	Sequoia	1
34	Haystack	7	79	Provenance-curious	3	124	Ghoshal et al. [118]	1	169	SPROV	1
35	Vansummeren et al. [119]	7	80	Arab et al.	3	125	Milieu	1	170	WASA	1
36	General purpose provenance library	7	81	YesWorkflow	3	126	Minemu	1			
37	ESSW	6	82	libdft	3	127	Dytan	1			
38	Galaxy	6	83	GProM	3	128	LBL	1			
39	Mondrian	6	84	EEPS	2	129	GridDB	1			
40	CMCS	5	85	ProvenanceJS	2	130	Knime	1			
41	RAMP	5	86	Sweave	2	131	PiCloud	1			
42	SPIDER	5	87	UV-CDAT	2	132	pomsets	1			
43	GenePattern	5	88	Oracle Total Recall (DB)	2	133	Giovanni	1			
44	Triana	5	89	Travos	2	134	Clario	1			
45	ReproZip	5	90	ATOM	2	135	Azkaban	1			

[51] Y. Zhao, M. Hategan, and et al, "Swift: Fast, Reliable, Loosely Coupled Parallel Computation," in *IEEE International Conference*

TABLE 14
Provenance Systems identified in the second review process and their appearances(I)

Ref.	System	Num.	Ref.	System	Num.	Ref.	System	Num.	Ref.	System	Num.
1	myGrid/ Taverna	93	26	PANDA	11	51	StarFlow	5	76	Pin	3
2	kepler	90	27	ReproZip	11	52	Triana	5	77	Pip	3
3	VisTrails	88	28	noWorkflow	10	53	Chiron	4	78	Ghoshal et al. [118]	3
4	PASS	73	29	Mondrian	9	54	Geo-Opera	4	79	Sahoo et al. [116]	3
5	Trio	41	30	Git2prov	8	55	IWBrowser	4	80	RDFProv	3
6	Chimera	29	31	ParaView	8	56	LabelFlow	4	81	RunMyCode.org	3
7	Karma	26	32	SAM	8	57	Marathe	4	82	SystemT	3
8	Cui 2000	25	33	Web Provenance	8	58	S-Language	4	83	Tariq et al. [120]	3
9	PASOA/PreServ	24	34	CDE	7	59	TDB	4	84	TREC	3
10	SPADE	21	35	Haystack	7	60	YesWorkflow	4	85	Tupelo	3
11	ZOOM	21	36	Janus	7	61	Arab et al.	4	86	WDO-It!	3
12	swift	18	37	OPUS	7	62	ARK	3	87	ZOO	3
13	Wings-Pegasus	18	38	Provmanager	7	63	Artemis	3	88	Gibson et al. [115]	2
14	ORCHESTRA	17	39	SOLE	7	64	Bochner et al. 2008	3	89	ATOM	2
15	Tioga	17	40	Vansummeren et al. [119]	7	65	DataHub	3	90	Bonsai	2
16	ES3	17	41	General purpose provenance library	6	66	DataTracker	3	91	Caravan	2
17	Lipstick	15	42	CMCS	5	67	DEEP	3	92	Condor	2
18	DBNotes	15	43	Galaxy	5	68	eScience Central	3	93	EEPS	2
19	COMAD	14	44	GenePattern	5	69	Flogger	3	94	ExSpan	2
20	Perm	14	45	Ipython	5	70	Knitr	3	95	Gaea	2
21	*Buneman* [62]	12	46	pPOD	5	71	libdft	3	96	Geodise	2
22	PLUS	12	47	ProvBench	5	72	LogicBlox	3	97	Geolineus	2
23	RAMP	12	48	Provenance Explorer	5	73	Milieu	3	98	GOOSE	2
24	Redux	12	49	SciCumulus	5	74	Oozie	3	99	Hadoopprov	2
25	Burrito	12	50	SPIDER	5	75	ourSpaces	3	100	IncPy	2

on Services Computing - Workshops (SCW'07), 2007, pp. 199–206.

[52] L. M. Gadelha Jr, B. Clifford, M. Mattoso *et al.*, "Provenance management in Swift," *Futur Gener Comput Syst*, vol. 27, no. 6, pp. 775–780, 2011.

[53] P. Groth, S. Miles, and L. Moreau, "PReServ: Provenance recording for services," *UK e-Science All Hands Meeting*, 2005.

[54] M. Stonebraker, J. Chen, N. Nathan *et al.*, "Tioga: Providing Data Management Support for Scientific Visualization Applications," in *Proc. of VLDB'93*, 1993, pp. 25–38.

[55] A. Woodruff and M. Stonebraker, "Supporting Fine-grained Data Lineage in a Database Visualization Environment," in *Proc. of ICDE'97*, 1997, pp. 91–102.

[56] A. Aiken, J. Chen, M. Stonebraker, and A. Woodruff, "Tioga-2: A Direct Manipulation Database Visualization Environment," in *Proc. of the Twelfth International Conference on Data Engineering (ICDE'96)*, 1996, pp. 208–217.

[57] E. Deelman, G. Singh, M.-H. Su *et al.*, "Pegasus: A framework for mapping complex scientific workflows onto distributed systems," *Scientific Programming*, vol. 13, no. 3, pp. 219–237, 2005.

[58] J. Kim, E. Deelman, Y. Gil *et al.*, "Provenance Trails in the Wings/Pegasus System," *Concurr Comput*, vol. 20, no. 5, pp. 587–597, 2008.

[59] E. Deelman, J. Blythe, Y. Gil *et al.*, "Pegasus: Mapping scientific workflows onto the grid," in *Proc. of the 2nd European Across Grids Conference (EAGC'04)*. Springer, 2004, pp. 11–20.

[60] Y. Gil, V. Ratnakar, E. Deelman *et al.*, "Wings for pegasus: Creating large-scale scientific applications using semantic representations of computational workflows," in *Proc. of AAAI'07*, vol. 22, no. 2, 2007, p. 1767.

[61] E. Deelman, G. Mehta, G. Singh *et al.*, "Pegasus: mapping large-scale workflows to distributed resources," in *Workflows for e-Science*. Springer, 2007, pp. 376–394.

[62] P. Buneman, A. Chapman, and J. Cheney, "Provenance management in curated databases," in *Proceedings of the ACM SIGMOD International Conference on Management of Data, Chicago, Illinois, USA, June 27-29, 2006*, 2006, pp. 539–550.

[63] A. Gehani and D. Tariq, "SPADE: support for provenance auditing in distributed environments," in *Proc. of the 13th International Middleware Conference*, 2012, pp. 101–120.

[64] A. Gehani and M. Kim, "Mendel: Efficiently verifying the lineage of data modified in multiple trust domains," in *Proc. of the 19th ACM International Symposium on High Performance Distributed Computing (HPDC'10)*. ACM, 2010, pp. 227–239.

[65] Z. G. Ives, N. Khandelwal, A. Kapur, and M. Cakir, "ORCHESTRA: Rapid, Collaborative Sharing of Dynamic Data," in *Proc. of CIDR'05*, 2005, pp. 107–118.

[66] T. J. Green, G. Karvounarakis, Z. G. Ives, and V. Tannen, "Update Exchange with Mappings and Provenance," in *Proc. of VLDB'07*, 2007, pp. 675–686.

[67] T. J. Green, G. Karvounarakis, N. E. Taylor *et al.*, "ORCHESTRA: facilitating collaborative data sharing," in *Proc. of MOD'07*, 2007, pp. 1131–1133.

[68] B. Glavic and G. Alonso, "Perm: Processing Provenance and Data on the same Data Model through Query Rewriting," in *Proc. of the 25th IEEE International Conference on Data Engineering (ICDE'09)*, 2009, pp. 174–185.

[69] ——, "Provenance for Nested Subqueries," in *Proc. of the 12th International Conference on Extending Database Technology (EDBT'09)*, 2009, pp. 982–993.

[70] ——, "The Perm Provenance Management System in Action," in *Proc. of MOD'09 (Demonstration Track)*, 2009, pp. 1055–1058.

[71] B. Arab, D. Gawlick, V. Radhakrishnan, H. Guo, and B. Glavic, "A generic provenance middleware for queries, updates, and transactions," in *6th Workshop on the Theory and Practice of Provenance,*

TABLE 15
Provenance Systems identified in the second review process and their appearances(II)

Ref.	System	Num.	Ref.	System	Num.	Ref.	System	Num.	Ref.	System	Num.
101	NetLogo	2	126	Aurora	1	151	GExpLine	1	176	Merlin	1
102	Oracle Total Recall (DB)	2	127	Azkaban	1	152	Giovanni	1	177	MetaML	1
103	PAPEL	2	128	BDBMS	1	153	GridDB	1	178	Windows Workflow Foundation	1
104	Prefuse	2	129	CAMERA	1	153	GridSpace	1	179	Minebench	1
105	ProTracer	2	130	Caps	1	155	gryPhyN	1	180	Minemu	1
106	ProvAbs	2	131	CAPTRA	1	156	Gspan and close-graph	1	181	ML	1
107	Provenance map orbiter	2	132	CARE	1	157	Hi-Fi	1	182	MMS	1
108	Provenancecurious	2	133	CARMEN	1	158	Hive	1	183	MPO	1
109	ProvenanceJS	2	134	Causeway	1	159	Hyperloglog	1	184	Myexperiment	1
110	Story Book	2	135	CCDB	1	160	IADB	1	185	NetTrails	1
111	Subzero	2	136	CenterTrack	1	161	Ibis	1	186	Odyssey-SCM	1
112	Sweave	2	137	Century	1	162	InsightNotes	1	187	Opql	1
113	Travos	2	138	Clario	1	163	IWBase	1	188	OSLO	1
114	TrustCloud	2	139	Collage	1	164	JANUS	1	189	Panorama	1
115	UV-CDAT	2	140	CRISTAL	1	165	Kairos	1	190	Pastry	1
116	VIEW	2	142	Cytoscape	1	166	Kieker	1	191	PeerTrust	1
117	ViNE	2	142	D3S	1	167	Knime	1	192	PiCloud	1
118	Vismashup	2	143	Dexy	1	168	LabBase	1	193	PlanetFlow	1
119	W7	2	144	Dremel	1	169	LBL	1	194	PMAF	1
120	Myers et al. [117]	1	145	Dytan	1	170	Log4j	1	195	POMELO	1
121	ACCESSPROV	1	146	ESP2Net	1	171	MACEDON	1	196	pomsets	1
122	Amazonia	1	147	FilmTrust	1	172	Mars	1	197	PRISM	1
123	Antfarm	1	148	Flashback	1	173	Matrioshka	1	198	Prizms	1
124	A-R-E	1	149	Flexiq	1	174	MCA	1	199	PROB	1
125	Ariadne	1	150	GATES	1	175	Mendel	1	200	Propeller	1

TABLE 16
Provenance Systems identified in the second review process and their appearances(III)

Ref.	System	Num.	Ref.	System	Num.	Ref.	System	Num.	Ref.	System	Num.
201	ProPub	1	214	ReTrace	1	227	Sig.ma	1	240	TIUPAM	1
202	Provenance Browser	1	215	ReVirt	1	228	Sourcetrac	1	241	Tracefs	1
203	ProvGen	1	216	S2Logger	1	229	SPROV	1	242	Tribeca	1
204	PROVglish	1	217	SAF	1	230	Spyglass	1	243	TrustMe	1
205	ProvStore	1	218	SAGA	1	231	StatJR	1	244	Unicorn	1
206	ProvToolbox	1	219	SciPhy	1	232	Subdue	1	245	Utopia	1
207	PubFlow	1	220	Scorpion	1	233	SUSE	1	246	Valgrind	1
208	PUG	1	221	SecureFlow	1	234	Tabulator	1	247	Vle-wfbus	1
209	Qfix	1	222	seL4	1	235	Talos	1	248	W RBAC	1
210	Quill++	1	223	SELinks	1	236	TAP	1	249	WASA	1
211	Rasdaman	1	224	Sequoia	1	237	TelegraphCQ	1	250	WiDS	1
212	RDataTracker	1	225	SHARE	1	238	Tiresias	1	251	Ymaldb	1
213	RecProv	1	226	Sieve	1	239	Titian	1			

TaPP'14, Cologne, Germany, June 12-13, 2014, 2014.

[72] X. Niu, R. Kapoor, B. Glavic *et al.*, "Interoperability for Provenance-aware Databases using PROV and JSON," in *Proc. of TaPP'15*, 2015.

[73] B. Glavic, R. J. Miller, and G. Alonso, "Using SQL for Efficient Generation and Querying of Provenance Information," in *In search of elegance in the theory and practice of computation: a Festschrift in honour of Peter Buneman*, 2013, pp. 291–320.

[74] B. Glavic, "Perm: Efficient Provenance Support for Relational Databases," Ph.D. dissertation, University of Zurich, 2010, available at: http://www.zora.uzh.ch/44573/1/dissGlavic.pdf. Last visited on August 2017.

[75] B. Arab, D. Gawlick, V. Krishnaswamy *et al.*, "Reenacting Transactions to Compute their ProvenA system prototype for warehouse view maintenanceance," Illinois Institute of Technology, Technical Report IIT/CS-DB-2014-02, 2014, available at: http://cs.iit.edu/ dbgroup/pdfpubls/AD14.pdf. Last visited on August 2017.

[76] S. Lee, Y. Tang, S. Köhler *et al.*, "An Efficient Implementation Of Game Provenance In DBMS," Illinois Institute of Technology, Technical Report IIT/CS-DB-2015-02, 2015.

[77] X. Niu, R. Kapoor, and B. Glavic, "Heuristic and Cost-based Optimization for Provenance Computation," in *Proc. of TaPP'15*, 2015.

[78] J. Frew and P. Slaughter, "Es3: A demonstration of transparent provenance for scientific computation," in *Proc. of IPAW'08*, ser. LNCS, vol. 5272. Springer, 2008, pp. 200–207.

[79] J. Frew, D. Metzger, and P. Slaughter, "Automatic capture and reconstruction of computational provenance," *Concurr Comput*, vol. 20, no. 5, pp. 485–496, 2008.

[80] S. Bowers, T. M. McPhillips, and B. Ludäscher, "Provenance in collection-oriented scientific workflows," *Concurr Comput*,

TABLE 17
A brief description of the systems

[S1] VisTrails	It is a scientific workflow system developed at the University of Utah. One remarkable aspect of this system is the notion of provenance of workflow evolution [26].
[S2] myGrid/Taverna	It is a domain-independent Workflow Management System. Initially created by the *myGrid* team, it is now an Apache Incubator project [27].
[S3] Kepler provenance	Kepler-provenance is an add-on module which adds provenance features to the Kepler Scientific Workflow System presented in [28].
[S4] PASS	PASS (Provenance Aware Storage System) was developed at Harvard University as a storage system that supports the automatic collection and maintenance of provenance [29].
[S5] Trio	Developed at Stanford was defined as a new kind of database management system which extends relational databases with support for provenance and uncertainty [32] [33] [34] [35] [36][37].
[S6] Karma	It is a provenance collection and management system developed at Indiana University. It collects provenance for data-centric workflows in a service oriented architecture [38][39].
[S7] Chimera	Chimera is a part of the *GryPhyN* project, a research project which develops techniques for processing and managing large distributed data sets in data grids [40].
[S8] ZOOM* UserViews	ZOOM, for short, presents a model of provenance for scientific workflows. Its main goal is to provide users with an interface to query provenance information provided by a workflow system as well as to help them construct an appropriate user view [43][45].
[S9] Cui 2000	We refer as *Cui 2000* to the lineage tracing system developed by Cui and Widom in 2000 [47][48][49], and which is based on the WHIPS [50] data warehousing system.
[S10] Swift	A successor to Chimera (an outgrowth of the Chimera's Virtual Data Language), Swift is a scalable and reliable Grid workflow system that bridges scientific workflows with parallel computing [51] [52].
[S11] PASOA/PreServ	PreServ (Provenance Recording for Services), developed within the PASOA project, is a software package that allows developers to integrate process documentation recording into their applications [53].
[S12] Tioga	Tioga is a DBMS-centric visualization tool developed at Berkeley [54] [55]. Based on this tool, Tioga-2 [56] was designed in 1996 as a more powerful and much easier to program direct manipulation system.
[S13] Wings-Pegasus	Wings-Pagasus constitute two different tools which have been integrated to form a provenance system which works as a whole used for grid computing [58] [60].
[S14] Buneman	In [62], Buneman et al. present a proposal in which the user's actions are tracked while browsing source databases and copying data into a target curated database, in order to record the user's actions in a convenient, queryable form. This proposal is known by the *copy-paste-model* . In this review, we refer to this proposal as *Buneman,* for short.
[S15] SPADE	SPADE was developed as a distributed service for collecting, certifying, and querying the provenance of Grid data [63].
[S16] ORCHESTA	It is a collaborative data sharing system which focuses mainly on managing disagreement among multiple data representations and instances [65].
[S17] Perm-GProM	GProM (Generic Provenance Middleware) is defined as a generic provenance database middleware that enables computation of provenance for queries, updates, and transactions over several database back-ends (e.g., Oracle) [71].
[S18] ES3	The Earth System Science Server (ES3) is a software system designed for automatically and transparently capturing, managing, and reconstructing the provenance of arbitrary, unmodified computational sequences [78].
[S19] COMAD	CoMaD (Collection-Oriented Modeling and Design) is a provenance annotation-based framework, implemented as a part of the Kepler Scientific Workflow System [80].
[S20] DBNotes	DBNotes is an annotation management system for relational database systems [82].
[S21] Lipstick	It is a framework that marries database-style and workflow provenance models, capturing internal state as well as fine-grained dependencies in workflow provenance [83].
[S22] Redux	REDUX is a provenance management system which generates workflow provenance automatically during runtime. It uses the Windows Workflow Foundation (WinWF) [84] as workflow engine.
[S23] BURRITO	It is a Linux-based system that helps programmers organize, annotate and recall past insights about their experiments [85] [86].
[S24] PLUS	PLUS is a provenance manager that was inspired by U.S. government requirements to enable provenance capture, storage and use across multi-organizational systems [112].
[S25] RAMP	RAMP (Reduce And Map Provenance) is an extension to Hadoop that supports provenance capture and tracing for MapReduce workflows. More specifically, *MapReduce* is a programming model and an associated implementation for processing and generating large data sets [114].

vol. 20, no. 5, pp. 519–529, 2008.

[81] D. Bhagwat, L. Chiticariu, W. C. Tan, and G. Vijayvargiya, "An annotation management system for relational databases," *The VLDB Journal*, vol. 14, no. 4, pp. 373–396, 2005.

[82] L. Chiticariu, W. C. Tan, and G. Vijayvargiya, "DBNotes: A Post-It System for Relational Databases based on provenance," in *Proc. of MOD'05*. ACM, 2005, pp. 942–944.

[83] Y. Amsterdamer, S. B. Davidson, D. Deutch *et al.*, "Putting Lipstick on Pig: Enabling Database-style Workflow Provenance," *PVLDB*, vol. 5, no. 4, pp. 346–357, 2011.

[84] R. S. Barga and L. A. Digiampietri, "Automatic capture and efficient storage of e-Science experiment provenance," *Concurr Comput*, vol. 20, no. 5, pp. 419–429, 2008.

[85] P. J. Guo and M. Seltzer, "BURRITO: Wrapping Your Lab Notebook in Computational Infrastructure," in *Proc. of TaPP'12*, 2012.

[86] P. J. Guo, "Software Tools to Facilitate Research Programming," Ph.D. dissertation, Stanford University, May 2012.

[87] A. Cuzzocrea, "Big data provenance: State-of-the-art analysis and emerging research challenges," in *Proc. of the Workshops of the EDBT/ICDT 2016*, 2016, pp. 797–800.

[88] G. Dogan, "A survey of provenance in wireless sensor networks," *Ad Hoc & Sensor Wireless Networks*, vol. 30, no. 1-2, pp. 21–45, 2016.

[89] C. Wang, W. Zheng, and E. Bertino, "Provenance for wireless sensor networks: A survey," *Data Science and Engineering*, vol. 1, no. 3, pp. 189–200, 2016.

[90] Y. S. Tan, R. K. L. Ko, and G. Holmes, "Security and data accountability in distributed systems: A provenance survey," in *Proc. of HPCC'13*. IEEE Computer Society, 2013, pp. 1571–1578.

[91] D. Gammack, S. Scott, and A. P. Chapman, "Modelling provenance collection points and their impact on provenance graphs," in *Proc. of IPAW'16*, 2016, pp. 146–157.

[92] F. Chirigati, J. Freire, D. Koop, and C. Silva, "Vistrails provenance traces for benchmarking," in *Proc. of the Joint EDBT/ICDT 2013 Workshops*, 2013, pp. 323–324.

[93] C. E. Scheidegger, H. T. Vo *et al.*, "Querying and re-using workflows with vistrails," in *Proc. of MOD'08*, 2008, pp. 1251–1254.

[94] P. Alper, K. Belhajjame, and C. A. Goble, "Static analysis of taverna workflows to predict provenance patterns," *Futur Gener Comput Syst*, vol. 75, pp. 310–329, 2017.

[95] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li, "Taverna: a tool for the composition and enactment of bioinformatics workflows," *Bioinformatics*, vol. 20, no. 17, pp. 3045–3054, 2004.

[96] P. Missier, S. Soiland-Reyes, S. Owen *et al.*, "Taverna, reloaded," in *Proc. of the International conference on scientific and statistical database management (SSDBM'10)*, 2010, pp. 471–481.

[97] D. Crawl and I. Altintas, "A provenance-based fault tolerance

mechanism for scientific workflows," in *Proc. of IPAW'08*, 2008, pp. 152–159.

[98] M. R. Huq, A. Wombacher, and P. M. Apers, "Inferring fine-grained data provenance in stream data processing: reduced storage cost, high accuracy," in *Proc. of DEXA'11*, 2011, pp. 118–127.

[99] Y. L. Simmhan, B. Plale, and D. Gannon, "Karma2: Provenance management for data-driven workflows," *Web Services Research for Emerging Applications: Discoveries and Trends: Discoveries and Trends*, vol. 317, 2010.

[100] A. Marinho, D. de Oliveira, E. Ogasawara *et al.*, "Deriving scientific workflows from algebraic experiment lines: A practical approach," *Futur Gener Comput Syst*, vol. 68, pp. 111–127, 2017.

[101] M. Wilde, M. Hategan, J. M. Wozniak *et al.*, "Swift: A language for distributed parallel scripting," *Parallel Computing*, vol. 37, no. 9, pp. 633–652, 2011.

[102] D. Garijo, Y. Gil, and O. Corcho, "Abstract, link, publish, exploit: An end to end framework for workflow sharing," *Futur Gener Comput Syst*, 2017.

[103] Y. Gil, V. Ratnakar, J. Kim *et al.*, "Wings: Intelligent workflow-based design of computational experiments," *IEEE Intelligent Systems*, vol. 26, no. 1, pp. 62–72, 2011.

[104] A. Gehani, H. Kazmi, and H. Irshad, "Scaling spade to big provenance," in *Proc. of TaPP'16*. USENIX Association, 2016, pp. 26–33.

[105] M. Stamatogiannakis, H. Kazmi, H. Sharif *et al.*, "Trade-offs in automatic provenance capture," in *Proc. of IPAW'16*, 2016, pp. 29–41.

[106] S. C. Chan, A. Gehani, J. Cheney *et al.*, "Expressiveness benchmarking for system-level provenance," in *Proc. of TaPP'17*. USENIX Association, 2017.

[107] T. J. Green and V. Tannen, "The semiring framework for database provenance," in *Proc. of PODS'17*. ACM, 2017, pp. 93–99.

[108] Z. G. Ives, T. J. Green, G. Karvounarakis *et al.*, "The orchestra collaborative data sharing system," *ACM SIGMOD Record*, vol. 37, no. 3, pp. 26–32, 2008.

[109] P. Macko and M. Seltzer, "Provenance map orbiter: Interactive exploration of large provenance graphs." in *Proc. of TaPP'11*, 2011.

[110] B. Blaustein, L. Seligman, M. Morse *et al.*, "Plus: Synthesizing privacy, lineage, uncertainty and security," in *Proc. of the 24th International Conference on Data Engineering Workshop (ICDEW'08)*, 2008, pp. 242–245.

[111] A. Chapman, M. D. Allen, B. Blaustein *et al.*, "Plus: Provenance for life, the universe and stuff," *VLDB'10, VLDB Endowment*, pp. 13–17, 2010.

[112] A. Chapman, B. T. Blaustein, L. Seligman, and M. D. Allen, "Plus: A provenance manager for integrated information," in *Proc. of the IEEE International Conference on Information Reuse and Integration (IRI'11)*. IEEE, 2011, pp. 269–275.

[113] R. Ikeda, H. Park, and J. Widom, "Provenance for generalized map and reduce workflows," in *Proc. of the Fifth Biennial Conference on Innovative Data Systems (CIDR'11)*, 2011, pp. 273–283.

[114] H. Park, R. Ikeda, and J. Widom, "Ramp: A system for capturing and tracing provenance in mapreduce workflows," *PVLDB*, vol. 4, no. 12, pp. 1351–1354, 2011.

[115] A. Gibson, M. Gamble, K. Wolstencroft, T. Oinn, C. A. Goble, K. Belhajjame, and P. Missier, "The data playground: An intuitive workflow specification environment," *Future Generation Comp. Syst.*, vol. 25, no. 4, pp. 453–459, 2009.

[116] S. S. Sahoo, R. S. Barga, J. Goldstein, and A. P. Sheth, "Provenance algebra and materialized view-based provenance management," in *Proceedings of the 2rid International Provenance and Annotation Workshop. Berlin: Springer*, 2008, pp. 531–540.

[117] J. D. Myers, L. Marini, P. Bajcsy, R. Kooper, Y. Liu, R. McGrath, J. Futrelle, T. McLaren, A. Collier, and A. Rodriguez, "A Digital Synthesis Framework for Virtual Observatories," *AGU Fall Meeting Abstracts*, Dec. 2008.

[118] D. Ghoshal and B. Plale, "Provenance from log files: a bigdata problem," in *Joint 2013 EDBT/ICDT Conferences, EDBT/ICDT '13, Genoa, Italy, March 22, 2013, Workshop Proceedings*, 2013, pp. 290–297.

[119] S. Vansummeren and J. Cheney, "Recording provenance for SQL queries and updates," *IEEE Data Eng. Bull.*, vol. 30, no. 4, pp. 29–37, 2007.

[120] M. A. D. Tariq and A. Gehani, "Towards automated collection of application-level data provenance," in *Proceedings of the 4th Workshop on the Theory and Practice of Provenance (TaPP12). USENIX Association*, 2012.