The Foundations for Provenance on the Web[1]

Luc Moreau

November 3, 2009

[1]Submitted to Foundations and Trends in Web Science.

Abstract

Provenance, i.e., the origin or source of something, is becoming an important concern, since it offers the means to verify data products, to infer their quality, to analyse the processes that led to them, and to decide whether they can be trusted. For instance, provenance enables the reproducibility of scientific results; provenance is necessary to track attribution and credit in curated databases; and, it is essential for reasoners to make trust judgements about the information they use over the Semantic Web.

As the Web allows information sharing, discovery, aggregation, filtering and flow in an unprecedented manner, it also becomes very difficult to identify, reliably, the original source that produced an information item on the Web. Since the emerging use of provenance in niche applications is undoubtedly demonstrating the benefits of provenance, we contend that provenance can and should reliably be tracked and exploited on the Web, and we survey the necessary foundations to achieve such a vision.

Using multiple data sources, we have compiled the largest bibliographical database on provenance so far. This large corpus allows us to analyse emerging trends in the research community. Specifically, using the CiteSpace tool, we identify clusters of papers that constitute research fronts, from which we derive characteristics that we use to structure our foundational framework for provenance on the Web. We note that such an endeavour requires a multi-disciplinary approach, since it requires contributions from many computer science sub-disciplines, but also other non-technical fields given the human challenge that is anticipated.

To develop our vision, it is necessary to provide a definition of provenance that applies to the Web context. Our conceptual definition of provenance is expressed in terms of processes, and is shown to generalise various definitions of provenance commonly encountered. Furthermore, by bringing realistic distributed systems assumptions, we refine our definition as a query over assertions made by processes.

Given that the majority of work on provenance has been undertaken by the database, workflow and e-science communities, we review some of their work, contrasting approaches, and focusing on important topics we believe to be crucial for bringing provenance to the Web, such as abstraction, collections, storage, queries, workflow evolution, semantics and activities involving human interactions.

However, provenance approaches developed in the context of databases and workflows essentially deal with closed systems. By that, we mean that workflow or database management systems are in full control of the data they manage, and track their provenance within their own scope, but not beyond. In the context of the Web, a broader approach is required by which chunks of provenance representation can be brought together to describe the provenance of information flowing across multiple systems. This is the specific purpose of the Open Provenance

Vision, which is an approach that consists of controlled vocabulary, serialization formats and interfaces that allow the provenance of individual systems to be expressed, connected in a coherent fashion, and queried seamlessly. In this context, the Open Provenance Model is an emerging community-driven representation of provenance, which has been actively used by some twenty teams to exchange provenance information according to the Open Provenance Vision.

Having identified an open approach and a model for provenance, we then look at techniques that have been proposed to expose provenance over the Web. We also study how Semantic Web technologies have been successfully exploited to express, query and reason over provenance. Symmetrically, we also identify how Semantic Web technologies such as RDF underpinning the Linked Data effort bring their own difficulties with respect to provenance.

A powerful argument for provenance is that it can help make systems transparent, so that it becomes possible to determine whether a particular use of information is appropriate under a set of rules. Such capability helps make systems and information accountable. To offer accountability, provenance itself must be authentic, and rely on security approaches that we review. We then discuss systems where provenance is the basis of an auditing mechanism to check past processes against rules or regulations. In practice, not all users want to check and audit provenance, instead, they may rely on measures of quality or trust; hence, we review emerging provenance-based approaches to compute trust and quality of data.

Contents

1 Introduction **3**
- 1.1 Drivers for Provenance . . . 3
- 1.2 Provenance for Web Science . . . 5
- 1.3 A Web Science View of Provenance . . . 6

2 Analysis of the Provenance Literature **7**
- 2.1 The Provenance Bibliography . . . 7
- 2.2 New Research Fronts . . . 9
- 2.3 Analysis of Research Trends . . . 10
- 2.4 Summary . . . 16

3 Definition of Provenance **18**
- 3.1 Dictionary Definition . . . 18
- 3.2 Definition of Provenance in Computer Systems . . . 19
- 3.3 Mashup Exemplar Application . . . 20
- 3.4 Alternative Definitions of Provenance . . . 21
- 3.5 Assumptions . . . 24
- 3.6 Provenance: a Query over Process Assertions . . . 26
- 3.7 Summary . . . 27

4 Provenance in Workflows and Databases **28**
- 4.1 Views and Abstraction . . . 30
- 4.2 Data Collections and Streams . . . 32
- 4.3 Efficient Storage of Provenance . . . 34
- 4.4 Querying Provenance . . . 35
- 4.5 Workflow Evolution . . . 37
- 4.6 Provenance Semantics . . . 38
- 4.7 Human-Driven Workflows . . . 39
- 4.8 Summary . . . 39

5 The Open Provenance Vision **41**
- 5.1 Provenance-Aware Monolithic Application . . . 42
- 5.2 Provenance Inter-Operability across Components . . . 43

5.3 The Provenance Challenge Series . . . 44
5.4 The Open Provenance Model . . . 45
5.5 Provenance in Open Systems . . . 47
5.6 Broadening the Scope of Provenance beyond Closed Systems . . . 49
5.7 Summary . . . 50

6 Provenance, the Web and the Semantic Web 52
6.1 Publishing Provenance on the Web . . . 52
6.2 Semantic Web Techniques for Provenance . . . 53
6.3 Provenance for RDF . . . 55
6.4 Knowledge and Web Provenance . . . 57
6.5 Summary . . . 58

7 Accountability 59
7.1 Provenance and Security . . . 60
7.1.1 Access Control . . . 61
7.1.2 Provenance Integrity . . . 62
7.1.3 Liability and Accountability for Provenance . . . 63
7.1.4 Sensitivity of Provenance Information . . . 64
7.2 Accountability . . . 64
7.3 Data Quality and Trust . . . 66
7.4 Summary . . . 68

8 Conclusion 69
8.1 The Benefits of Provenance on the Web . . . 70
8.2 Future Research . . . 71

Chapter 1

Introduction

Provenance, i.e., the origin or source of something, is becoming an important concern for several research communities in computer science, since it offers the means to verify data products, to infer their quality, to analyse the processes that led to them, and to decide whether they can be trusted. In fact, provenance is an intrinsic property of data, which gives data value, when accurately captured. To motivate the need for provenance, we first review its potential benefits in several contexts: e-science, curated databases and Semantic Web. Furthermore, we show that the provenance philosophy is not restricted to data in computer systems, but applies also to real-life artifacts, such as ingredients in the food industry, parts in manufacturing and works of art (Section 1.1). Building on theoretical and practical results related to provenance, we propose a new, multi-disciplinary perspective of provenance, so that it can be developed on the Web (Section 1.2). We then outline our bibliography-based methodology to identify trends in the provenance research community, which we then use to structure our vision of provenance on the Web (Section 1.3).

1.1 Drivers for Provenance

As the e-science vision becomes reality [443, 444], researchers in the scientific community are increasingly perceived as providers of online data, which take the form of raw data sets from sensors and instruments, data products produced by workflow-based intensive computations [174], or databases resulting from sophisticated curation [53]. While science is becoming computation and data intensive, the fundamental tenet of the scientific method remains unchanged: experimental results need to be reproduceable. In contrast to a workflow, which can be viewed as a recipe that can be applied in the future, *provenance* is regarded as the equivalent of a logbook, capturing all the necessary steps involved in the actual derivation of a result, and which can be used to replay the execution that led to that result so as to validate it.

Curated databases typically represent the efforts of a dedicated group of people to produce a definitive description of some subject area [53]. They cover a vast range of application domains from Swiss-Prot[1], which is a protein knowledgebase that is manually annotated and reviewed, to Wikipedia[2], a crowd-sourced encyclopaedia, with increasingly sophisticated editorial processes. Such databases are generally published on the Web; they are heavily cross-referenced with, and include data from, other databases. Curated databases therefore bring out the problem of attribution (who initially created a description), and raise questions about the source, or *provenance*, of such descriptions (where were descriptions initially published).

Meanwhile, the Web has evolved into a network of blogs, information portals, and social bookmarking services which provide automated feeds between subscribers. As the Web allows information sharing, discovery, aggregation, filtering and flow in an unprecedented manner, it also becomes very difficult to identify, reliably, the original source that produced an information item on the Web. Without knowing the provenance of information, information services may not be able to undertake the necessary due diligence about their content, they may be the subject of fraud or spam, and overall they may be judged as unreliable. Provenance is in fact identified as one of the many salient factors that affect how users determine trust in content provided by Web information sources [442]. This view is echoed by Lynch [257], who argues that among the consequences of this shift to new highly distributed dissemination systems, will be a new emphasis on the provenance of data and metadata, and the need for information retrieval systems to permit users to factor in trust preferences about this information.

Where the Web originally consisted largely of documents intended to be read by humans, the Semantic Web [430] envisions a Web of information and knowledge processable by computer systems which undertake automated reasoning. Central to this effort are RDF [446] and OWL [437], the frameworks in which to express metadata, vocabularies and perform associated reasoning. This vision is being deployed by means of Linked Data [431, 452], an information space in which data is being enriched by typed links expressed in RDF [446], cross-referencing data sets, in a machine-processable fashion. Given the possibility for anybody (or system) to publish sets of Linked Data that refer to others, reasoners will need explicit representations of provenance information in order to make trust judgements about the information they use [426].

The issue of provenance is in no way limited to data, information or knowledge. It also applies to physical artifacts, for example in the food industry. From wine to meat, from dairy products to whisky, from coffee to vegetables, the food industry is very keen to be able to demonstrate the origin of the ingredients we purchase and eat. Understanding the provenance of food, i.e. its origin, how it

[1] www.uniprot.org
[2] www.wikipedia.org

is produced, transported, and delivered to us, is turned into a competitive advantage by the food industry, since it allows it to demonstrate quality (in taste, in carbon footprint, or in ethics). Furthermore, across the world, governments and associated regulatory authorities are interested in food safety, and typically require the traceability of food. Likewise, manufacturers focus on compliance and traceability initiatives for a variety of reasons. Understanding past processes is critical to discover bottlenecks, inefficiencies, wastage, and learn how to improve them. Exact traceability is essential to manage product recalls efficiently and minimise their economic impact. Similarly to the food industry, provenance of products is used to build customer trust. And of course, in the context of art, the provenance of art objects is so important that available evidence is typically produced before auctions in order to maximize the price obtained for these objects.

1.2 Provenance for Web Science

Web science is the emerging interdisciplinary field that aims to understand the Web, engineer its future and ensure its social benefit [429]. In the context of Web science, trust is recognised as one of the important concerns associated with the Web [181]: there is a broad consensus that trust in content could be derived if the transformations and derivations that resulted in such content can be known. Hence, given that the Web currently provides little support for provenance, the topic of provenance is becoming recognized as an important subject of investigation [452] in this context.

Like Web science, there is a multi-disciplinary facet to provenance. First, within computer science, multiple sub-disciplines are involved including database, systems, escience, grid, Semantic Web, and security. Second, provenance can be exploited to provide new services to the scientific community, businesses, and all Web users. It has the potential to make systems more transparent, and therefore auditable. As a result, it is a strong contender technology to underpin information accountability [402]. While it can be used to perform compliance checks (such as conformance to process or checking that terms of data licensing are met), it also raises issues related to privacy. Thus, societal, legal, and business perspectives on provenance could potentially have a wide impact on its use on the Web.

Our aim in this document is to survey the technical aspects of provenance that are relevant to Web science, but also to draw attention to the potential multi-disciplinary opportunities that they bring. Provenance, as a technical subject of study, is by no means a green field. The oldest publications discussed in this survey dates back to the late eighties. Importantly, the interest of provenance has been growing dramatically, as illustrated by the number of publications on the topic (see Figure 2.1, page 8, to be discussed in the next chapter). We have identified over 400 publications on provenance, 200 of which have been published

over the last two years.

Several surveys already exist, but, to some extent, work has so far been broadly surveyed on a per discipline basis, as illustrated by Simmhan *et al.*'s review of provenance in e-Science [360], Bose and Frew's survey of provenance for scientific processing [36], Cheney *et al.*'s survey of database provenance [226], and Glavic and Dittrich's classification of approaches [178]. In this article, we aim to break such silos, and try to investigate cross-cutting concerns that are relevant to providing provenance of information on the Web.

In fact, society is now at a turning point since it is presented with a unique opportunity, which will require social and technical changes: it is the author's belief that *society can and should reliably track and exploit the provenance of information on the Web*. To achieve this vision, the research output from all disciplines investigating provenance should be integrated into a coherent approach, for which we propose a foundational framework here. For instance, the work undertaken by the workflow community on provenance is very relevant to the flow of information that we now observe on the Web, since it can help track provenance as information flows through distributed services. Given that much of the data available on the Web is actually stored in databases, provenance research in the context of databases is essential, since it tracks provenance as data changes within databases. The work focusing on making provenance secure and non forgeable is also relevant to our goal of reliably tracking information on the Web.

1.3 A Web Science View of Provenance

Having compiled the most extensive bibliography on provenance so far, Web science provides us with tools and techniques to analyse this research topic. Simple metrics such as citation count can help us identify the most popular papers. However, citation analysis can help us obtain a deeper insight in the different subfields of this subject of study. Using clustering techniques, we can identify emerging research fronts dealing with different concerns; using tag clouds [427], we can summarize these concerns in a visual manner.

This survey is structured as follows. In Chapter 2, we undertake an analysis of the provenance literature, we discuss key topics of interest and we identify landmark papers. In Chapter 3, we discuss a broad definition of provenance that would apply to the Web and compare it with various alternative definitions that we recast in a Web context. Provenance has traditionally been studied in the context of workflows and databases, and we contrast the work undertaken in these communities (Chapter 4). By some measures, these approaches can be regarded as closed; hence, in Chapter 5, we introduce a vision of provenance for open systems such as the Web. We then discuss in Chapter 6 issues of provenance related to Web technologies. We finally tackle the problem of accountability in Chapter 7, before some concluding remarks in Chapter 8.

Chapter 2

Analysis of the Provenance Literature

Using multiple data sources, we have compiled the largest bibliographical database on provenance (in a technical sense) so far. The database[1] is made available online by the journal publisher [2]. This reasonably large corpus allows us to identify landmark papers and observe an acceleration of activities (Section 2.1), but more importantly, it allows us to analyse emerging trends in the research community. Specifically, using the tool CiteSpace [434, 435], we discover clusters of papers that constitute research fronts (Section 2.2). Six clusters have been identified and positioned in time, covering topics as varied as database, workflows, eScience, "Provenance Challenge", Open Provenance Model, Semantic Web and electronic notebooks. We then identify the key characteristics of these research fronts, and use them to structure our foundational framework for provenance on the Web, as well as the rest of the paper (Section 2.3).

2.1 The Provenance Bibliography

The provenance bibliographical database was compiled using multiple sources: the author's own original database, the ACM, IEEE, and Springer digital libraries, the DBLP computer science bibliography[3], and some programmes of provenance-specific events such as the International Provenance and Annotations workshops (IPAW'06, IPAW'08), and the Workshop on Theory and Practice of Provenance (TAPP'09). For each publication, we maintain the explicit list of publications it cites and its abstract. We applied the transitive closure and en-

[1] Note that this article's bibliography is divided in two sections: the first part consists of the provenance bibliography, whereas the second part starting page 125 refers to papers that do not have provenance as a specific focus.

[2] www.nowpublishers.com/web/ *to complete by publisher*

[3] http://dblp.uni-trier.de/

sured that each cited paper that contained the words 'provenance' or 'lineage' in its title was included in the database (provided this was a Computer Science paper). When a technical report was superseded by a published paper, the latter was preferred (and we ensured the published paper was assigned all its citations).

Figure 2.1 contains a histogram displaying the number of publications on provenance per year. A total of 425 papers have been identified. The first publication dates back from 1986 [22] and describes an auditing technique to assist analysts in understanding and validating data results. The histogram shows a definite trend in the research activity related to provenance, with about half the papers published in the last two years.

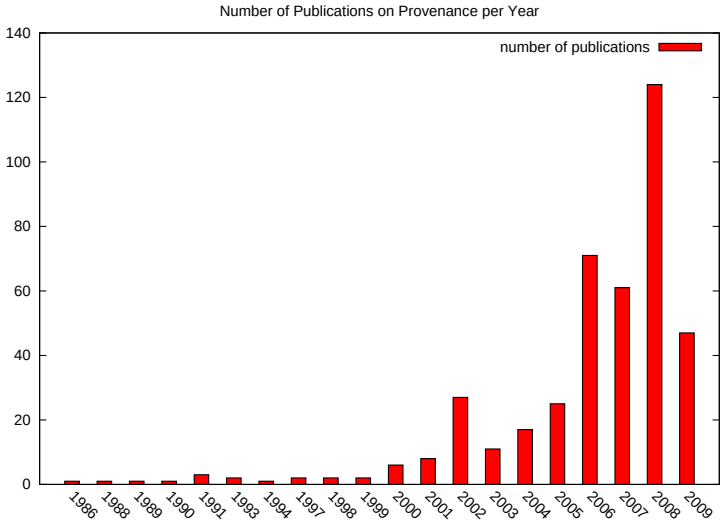


Figure 2.1: Number of Provenance Publications

In fact, several publication peaks can be observed in Figure 2.1; they coincide with events organised by the "provenance community" itself: in 2002, the first provenance workshop organized by Foster and Buneman; in 2006, the International Provenance and Annotation Workshop[4] (IPAW) workshop organised by Foster and Moreau; and, in 2008, the second IPAW workshop organised by Freire and Moreau and the first Provenance Challenge special issue edited by Moreau and Ludaescher [297].

From the provenance bibliography database, we have produced the list of the most cited papers (within the database). At the top of the list, we find Buneman *et al.*'s seminal paper on Where and Why provenance [53]. Two surveys [360, 36] appear in the top five. We note that this analysis does not reflect the impact of a publication outside the provenance community. For instance,

[4] www.ipaw.info

Buneman *et al.*'s paper attracts 387 citations in Google Scholar, but is cited only 116 times in this database.

Rank	Citations	Paper (first author:venue)		Rank	Citations	Paper (first author:venue)	
1	(116)	Buneman:ICDT01	[59]	26	(27)	Russell:tech02	[339]
2	(97)	Simmhan:SIGMOD05	[360]	27	(27)	Frew:SSDBM01	[157]
3	(65)	Foster:SSDBM02	[146]	28	(27)	Bowers:IPAW06	[39]
4	(62)	Cui:TODS00	[107]	29	(24)	Simmhan:ICWS06	[361]
5	(56)	Bose:ACMCS05	[36]	30	(24)	Moreau:OPM1.00	[292]
6	(49)	Moreau:CCPE08	[298]	31	(24)	Benjelloun:VLDB06	[25]
7	(47)	Woodruff:ICDE97	[406]	32	(24)	Buneman:FSTTCS2000	[58]
8	(47)	Miles:JOGC07	[279]	33	(23)	Barga:CCPE08	[17]
9	(44)	Groth:D3.1.1	[195]	34	(23)	Greenwood:AHM03	[189]
10	(40)	Bhagwat:VLDB04	[28]	35	(23)	Zhao:SWT03	[415]
11	(39)	Muniswamy-Reddy:USENIX06	[301]	36	(20)	Wang:VLDB90	[398]
12	(38)	Widom:CIDR05	[403]	37	(19)	Moreau:CACM08	[294]
13	(38)	Buneman:SIGMOD06	[51]	38	(18)	Buneman:ICDT07	[54]
14	(37)	Freire:IPAW06	[156]	39	(18)	Lanter:CGIS91	[242]
15	(36)	Groth:OPODIS04	[197]	40	(18)	Bavoil:VC05	[20]
16	(36)	Cui:VLDB03	[106]	41	(18)	Bowers:CCPE08	[41]
17	(33)	Fosterb:PROV02	[145]	42	(18)	Simmhan:IPAW06	[364]
18	(31)	Zhao:CCPE08	[419]	43	(17)	Miles:CCPE08	[282]
19	(31)	Groth:HPDC05	[198]	44	(17)	Zhao:IPAW06	[424]
20	(30)	Szomszor:ODBASE03	[375]	45	(17)	Braun:IPAW06	[44]
21	(30)	Cui:ICDE00	[104]	46	(17)	Myers:SWT03	[306]
22	(29)	Zhao:ISWC04	[421]	47	(16)	Kim:CCPE08	[232]
23	(29)	Altintas:IPAW06	[5]	48	(16)	Zhao:ICSNW04	[416]
24	(28)	Moreau:IPAW06	[291]	49	(16)	Frew:CCPE08	[159]
25	(27)	Green:PODS07	[188]	50	(16)	Tan:DBBUL07	[380]

Figure 2.2: Fifty Most Cited Publications

2.2 New Research Fronts

Emergent trends and abrupt changes in the scientific literature can be associated with internal as well as external causes [434]: typical internal causes include new discoveries and scientific breakthroughs; external ones may provoke scientists to study a matter from new perspectives. In the case of provenance, we conjecture that the development of the Grid as a technology for running scientific applications [440] and the UK e-science programme [444] have been two significant external triggering factors that have caused increasing number of researchers to focus on the provenance problem. Our conjecture is based on an analysis of the participants to the first provenance workshop and IPAW'08, who were predominantly from a grid computing and e-science background.

Chen [434] defines notions of research front and intellectual base.

> The concept of a *research front* was originally introduced by Price to characterize the transient nature of a research field. Price observed what he called the immediacy factor: There seems to be a tendency

> for scientists to cite the most recently published articles. In a given field, a research front refers to the body of articles that scientists actively cite.
>
> The concept of an *intellectual base* is useful to further clarify the nature of a research front. If we define a research front as the state of the art of a specialty (i.e., a line of research), what is cited by the research front forms its intellectual base.

To investigate provenance research fronts, we used Chen's citation analysis tool CiteSpace[5], and its definitions:

- Research front: Emerging thematic trends and surges of new topics;
- Intellectual base: Co-citation network;
- Cluster: Hybrid networks of co-cited articles and terms citing these articles;
- Labeling: Terms from titles, abstract and descriptions of abrupt frequency increase.

2.3 Analysis of Research Trends

Using CiteSpace's default configuration, we considered the bibliography's 1995 to 2009 time period, sliced in 1-yearly slices. For each slice, we consider a maximum of the 30 most cited publications. We also excluded the two most cited publications [59, 360] from our analysis, since their frequent citations resulted in a smaller, and not discriminating set of clusters. In this section, we analyse and provide our interpretation of the clustering produced by CiteSpace.

Six clusters have been identified by CiteSpace. They are graphically displayed in Figure 2.3. We have named them as follows:

Cluster 0: Security

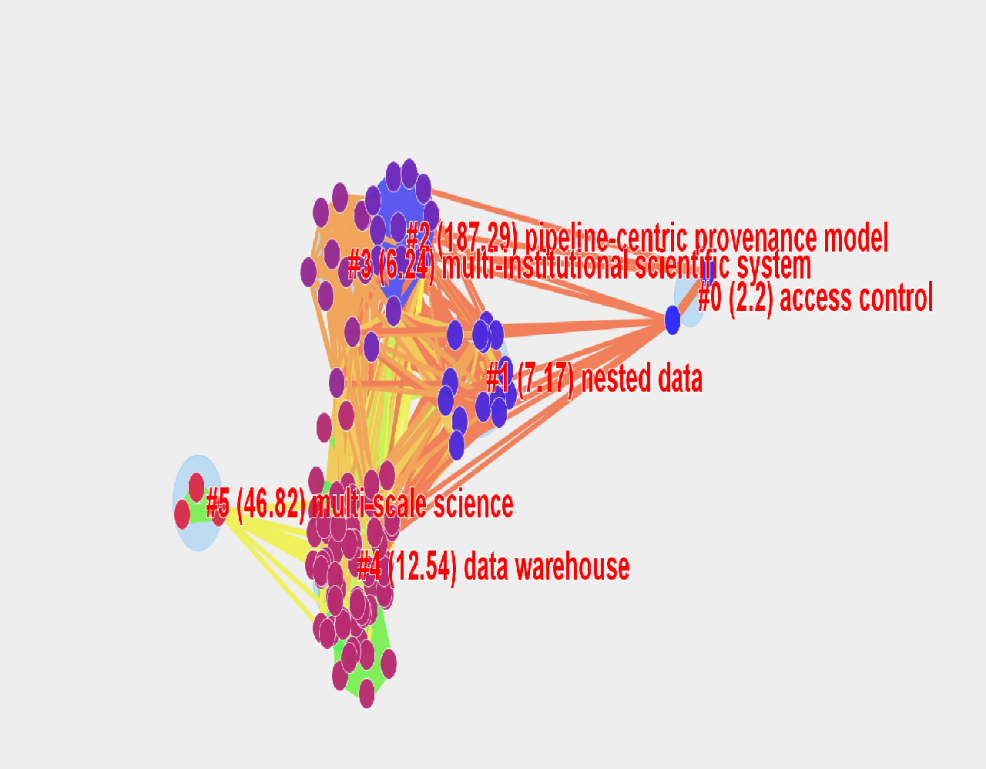
Cluster 1: Database 05–08, RDF and Open Provenance Model

Cluster 2: Workflow 05–08 and Database 08–09

Cluster 3: Provenance Challenge

Cluster 4: eScience and Database 02–05

Cluster 5: Electronic notebook



Cluster name:	First cluster label (See Figures 2.6 and 2.7)
Cluster 0: Security	(2.2) access control
Cluster 1: Database 05-08 and OPM	(7.17) nested data
Cluster 2: Workflow 05-08 and Database 06-08	(187.29) pipeline-centric provenance model
Cluster 3: Provenance Challenge	(6.24) multi-institutional scientific system
Cluster 4: eScience and DB 02-05	(12.54) data warehouse
Cluster 5: Electronic Logbook 02	(46.82) multi-scale science

Figure 2.3: Six Clusters in the Provenance Literature

Figures 2.4 and 2.5 contain tables summarising the intellectual base of these clusters, i.e., the papers cited by the research front; the tables are structured as follows. The first column identifies the cluster number. The second column is concerned with bursts, defined as follows: a burst is a set of fast-rising terms used by scientists in their latest publications. The second column contains the burst factor, which gives a measure of how often a paper is cited in the context of a burst in the citing paper; the higher the number, the more the publication is cited in papers with these fast-rising terms. The third column contains the betweenness centrality (it quantifies the importance of the node's position in a cluster): the higher the value, the more representative of the cluster. Sigma combines betweenness centrality and the burst rate to provide an indication of the

[5]http://cluster.cis.drexel.edu/∼cchen/citespace/

transformative strength of a publication in a given network over a time interval [435]; again, the higher the number, the stronger its transformative potential. We then find the publication year, the author and publication venue, and the bibliographic reference.

The cluster description is complemented by Figures 2.6 and 2.7 which contain the cluster labels and the research front. The research front consists of the citers to a cluster, whereas the cluster labels are obtained from salient features selected from the titles and abstracts of the citers; labels are combined with descriptors of abrupt frequency increase.

We remind the reader that this set of clusters is not a partition of the complete bibliography, since only the 30 most cited publications are selected for each year. Given the histogram of Figure 2.1, this means that a substantial number of publications from 2006–2009 do not appear in Figures 2.4 to 2.7. We note however that this is not a concern since our aim is to identify trends, rather than exhaustively categorize papers. The figures are provided for completeness, in order to guide readers who wish to investigate these clusters. We now describe each of them.

Cluster 0: Security Cluster 0 is the smallest and consists of two papers in its intellectual base [47, 377], identifying key issues related to provenance and security. The research front is formed by publications dating from 2007–2009, discussing issues pertaining to provenance access control, provenance authenticity and scalable security.

The following two clusters identify research trends in the database and workflow communities.

Cluster 1: Database 05–08, RDF and OPM This cluster's intellectual base consists of a set of publications from the database community, essentially investigating theoretical aspects of provenance [51, 188, 55, 89, 54, 25, 92, 53], as well as a survey [380], which shows a high citation burst, and two tutorials [62, 117]. In this cluster, we also find the publications about the Open Provenance Model [292, 299] – a community-driven provenance model —, the former having a high burst of citation, and a frequently cited paper about provenance in RDF [66]. The research front is predominantly composed of papers from the database community. Key topics discussed include nested data collection, dependency analysis and database technology.

Cluster 2: Workflow 05–08 and Database 06–08 Cluster 2 is the dual of Cluster 1. Its intellectual base consists of a vast majority of workflow papers [250, 174, 201, 97, 256, 294, 419, 39], a provenance-aware operating sytem [301], and a survey [36]; they are accompanied by database papers, which are also

Cluster	Burst	Centrality	Sigma	Year	Author:Venue	Citation
Cluster 0: Security (2/2)						
0		0.01	0	2008	Braun:HOTSEC08	[47]
0		0	0	2006	Tan:IPAW06	[377]
Cluster 1: Database 05-08, RDF and OPM (14/14)						
1	6.14	0.01	0.06	2007	Moreau:OPM1.00	[292]
1	3.25	0	0.03	2007	Tan:DBBUL07	[380]
1		0.03	0	2006	Buneman:SIGMOD06	[51]
1		0.01	0	2007	Green:PODS07	[188]
1		0.01	0	2005	Carroll:WWW05	[66]
1		0	0	2008	Buneman:TODS08	[55]
1		0	0	2007	Cheney:DBPL07	[89]
1		0	0	2007	Buneman:ICDT07	[54]
1		0	0	2006	Benjelloun:VLDB06	[25]
1		0	0	2006	Chiticariu:VLDB06	[92]
1		0	0	2008	Buneman:PODS08	[53]
1		0	0	2008	Davidson-Freire:SIGMOD08	[117]
1		0	0	2007	Buneman:SIGMOD07	[62]
1		0	0	2008	Moreau:OPM1.01	[299]
Cluster 2: Workflow 05-08 and Database 06-08 (13/13)						
2	4.99	0	0.01	2009	Levine:DFRWS09	[250]
2	4.34	0	0.01	2007	Gil-Deelman:IEEE07	[174]
2	4.34	0	0.01	2006	Agrawal:VLDB06	[1]
2	4.2	0	0.04	2008	Heinis:SIGMOD08	[216]
2	4.1	0	0.01	2009	Groth:TOIT09	[201]
2	3.4	0.03	0.18	2008	Chapman:SIGMOD08	[75]
2	3.4	0.02	0.14	2008	Clifford:CCPE08	[97]
2	3.4	0.01	0.11	2008	Ludaescher:CCPE08	[256]
2	3	0.02	0.17	2008	Moreau:CACM08	[294]
2		0.19	0	2005	Bose:ACMCS05	[36]
2		0.08	0	2006	Muniswamy-Reddy:USENIX06	[301]
2		0.04	0	2008	Zhao:CCPE08	[419]
2		0.02	0	2006	Bowers:IPAW06	[39]
Cluster 3: Provenance Challenge (11/11)						
3	9.4	0.09	0.19	2008	Moreau:CCPE08	[298]
3	4.33	0.02	0.13	2008	Barga:CCPE08	[17]
3	4.31	0	0.04	2008	Kim:CCPE08	[232]
3	3.31	0	0.06	2008	Miles:CCPE08	[282]
3	3.29	0	0	2008	Holland:CCPE08	[218]
3	2.8	0	0.02	2008	Golbeck:CCPE08	[182]
3		0.03	0	2006	Altintas:IPAW06	[5]
3		0.01	0	2008	Cohen-Boulakia:CCPE08	[99]
3		0.01	0	2008	Frew:CCPE08	[159]
3		0	0	2008	Bowers:CCPE08	[41]
3		0	0	2005	Bavoil:VC05	[20]

Figure 2.4: Clusters 0, 1, 2 and 3: Intellectual Base

Cluster	Burst	Centrality	Sigma	Year	Author:Venue	Citation
Cluster 4: eScience and DB 02-05 (22/59)						
4	6.3	0.03	0.15	2003	Greenwood:AHM03	[189]
4	6.11	0.15	0.31	1997	Woodruff:ICDE97	[406]
4	5.95	0.03	0.13	2004	Groth:OPODIS04	[197]
4	5.06	0.03	0.16	2000	Cui:ICDE00	[104]
4	4.89	0.05	0.19	2003	Zhao:SWT03	[415]
4	4.57	0.29	0.48	2000	Cui:ICDE00	[104]
4	3.91	0.04	0.19	2002	Foster:PROV02	[144]
4	3.76	0.16	0.4	2002	Foster:SSDBM02	[146]
4	3.58	0.01	0.08	2002	Bose:SSDBM02	[33]
4	3.41	0.04	0.22	2003	Szomszor:ODBASE03	[375]
4	3.28	0.03	0.19	2001	Frew:SSDBM01	[157]
4	3.19	0.02	0.17	2006	Groth:D3.1.1	[195]
4	3.16	0.04	0.21	2002	Goble:PROV02	[179]
4	2.77	0.01	0.12	2000	Buneman:FSTTCS2000	[58]
4		0.08	0	2002	Fosterb:PROV02	[145]
4		0.08	0	2005	Widom:CIDR05	[403]
4		0.03	0	2007	Miles:JOGC07	[279]
4		0.03	0	2003	Cui:VLDB03	[106]
4		0.03	0	2004	Bhagwat:VLDB04	[28]
4		0.02	0	2006	Freire:IPAW06	[156]
4		0.02	0	2005	Groth:HPDC05	[198]
4		0.02	0	2003	Myers:SWT03	[306]
Cluster 5: Electronic Logbook 02 (3/3)						
5		0.08	0	2003	Myers:CISE03	[309]
5		0	0	2002	Myers:PROV02	[307]
5		0	0	2002	Pancerella:PROV02	[315]

Figure 2.5: Clusters 4 and 5: Intellectual Base

concerned with systems [1, 216, 75]. The research front also consists of papers that, by and large, are system-oriented. We note that Semantic Web technology is quite a common thread in this cluster.

Cluster 3: Provenance Challenge In its intellectual base, Cluster 3 consists of papers published after the first provenance challenge [298], an inter-operability exercise between provenance systems. Its research front broadly consists of papers concerned with practical considerations for provenance.

Cluster 4: eScience and Database 02–05 Cluster 4's intellectual base is concerned with both eScience/grid and database research, with the following central papers: Foster *et al.*'s Chimera system [146], Woodruff and Stonebraker's fine-grained lineage [406], and Cui and Widom's lineage in data warehouses [104]. This period, which coincides with the first peak in the histogram of Figure 2.1, was very active, as illustrated by the cluster size of 59. In this cluster, several papers have significant publication burst, and have attracted substantial citations as illustrated by Figure 2.2.

Cluster 5: Electronic Logbook The last cluster consists of papers that were early advocate of nascent semantic web technologies, electronic logbook, and

Cluster 0: Security (size: 2)

(2.2) access control; (1.79) grouping provenance information; (1.79) fake picasso; (1.79) provenance security; (1.79) preventing history forgery; (1.79) new soa data-provenance framework; (1.1) provenance information; (1.1) scalable access control; (0.69) multi-institutional scientific system; (0.69) secure provenance

Research front:

Chong:TAPP09 [94], daCruz:CS09 [112], Groth:thesis07 [206], Hasan:FAST09 [215], Rosenthal:TAPP09 [338], Syalim:ISA09 [374], Tsai:ISADS07 [388], Tsai:SOCA07 [387]

Cluster 1: database 05-08 and OPM (size: 14)

(7.17) nested data; (7.17) dependency analysis; (7.17) sql query; (7.17) recording provenance; (5.38) data synchronization; (3.58) provenance semiring; (3.58) efficient provenance storage; (3.58) exploring scientific workflow provenance; (3.58) using hybrid query; (3.58) nested data collection; (3.58) lineage graph; (2.43) scientific workflow; (2.43) data provenance

Research front:

Anand:EDBT09 [7], Anand:SSDBM09 [8], Biton:VLDB07 [29], Buneman:ICDT07 [54], Buneman:IPAW06 [52], Buneman:SIGMOD07 [62], Chapman:DBBUL07 [73], Chebotko:escience08 [77], Cheney:DBBUL07 [85], Cheney:DBPL07 [89], Cheney:PLANX09 [87], Cheney:TRENDDB09 [226], daCruz:CS09 [112], Davidson:DBBUL07 [116], Ding:SS05 [124], Ding:tech05 [123], Factor:TAPP09 [135], Foster:DBBUL07 [148], Gibson:TAPP09 [172], Glavic:BTW07 [178], Green:PODS07 [188], Green:VLDB07 [187], Groth:ESAW09 [199], Hartig:LDOW09 [212], Hasan:SSS07 [214], Moreau:FOPM09 [296], Mutsuzaki:CIDE07 [305], Rosenthal:TAPP09 [338], Sarma:tech07 [347], Silles:USER09 [355], Sun:SIGMOD09 [373], Tan:DBBUL07 [380], Vansummeren:DBBUL07 [389]

Cluster 2: Workflow 05-08 and Database 06-08 (size: 13)

(187.29) pipeline-centric provenance model; (8.32) provenance model; (1.79) semantic annotation; (1.62) scientific workflow; (1.1) knowledge provenance; (1.1) collection-oriented scientific workflow run; (1.1) conceptual model

Research front:

Balis:eScience07 [14], Bowers:DILS07 [40], Buneman:SIGMOD07 [62], Chebotko:escience08 [77], Davidson:DBBUL07 [116], Groth:thesis07 [206], Groth:WORKS09 [194], Rio:GEOS07 [337], Wang:ICCS07 [395], ZhaoJ:thesis07 [417]

Cluster 3: Provenance Challenge (size: 11)

(6.24) multi-institutional scientific system; (4.85) provenance framework; (4.39) conceptual model; (2.77) e-science provenance; (2.43) scientific workflow; (2.2) collection-oriented scientific workflow run; (2.2) scientific data; (2.2) managing data provenance; (2.2) connecting scientific data; (2.2) project history; (1.79) tracking file; (1.79) kepler provenance framework; (1.79) provenance support; (1.79) creating visualization; (1.39) building practical provenance system; (1.1) constructing scientific publication package; (1.1) graphical interface; (1.1) using provenance; (1.1) provenance explorer-a;

Research front:

Bowers:DILS07 [40], Chapman:DBBUL07 [73], Chebotko:escience08 [77], daCruz:CS09 [112], Davidson:DBBUL07 [116], Groth:thesis07 [206], Hunter:IJDL07 [224], Miles:eScience07 [278], Mouallem:SSDBM09 [300], Scheidegger:TVCG07 [349], Simmhan:thesis07 [358], Stevens:BBIO07 [371], Wang:FGCS09 [394], ZhaoJ:thesis07 [417]

Figure 2.6: Labels and Research Front for Clusters 0 to 3

Cluster 4: eScience and DB 02-05 (size: 59)
(12.54) data warehouse; (12.08) scientific data; (8.96) virtual data language; (8.96) provenance recording; (8.96) warehousing environment; (8.96) practical lineage; (8.96) data grid environment; (8.96) view data; (8.92) data provenance; (7.17) managing scientific data lineage; (7.17) propagation module; (7.17) lineage retrieval; (7.17) tracing data lineage; (7.17) data provenance problem; (7.17) scientific data processing; (7.17) contemplating vision; (7.17) conceptual framework; (7.17) using schema transformation pathway; (7.17) provenance-aware sensor data storage; (7.17) realizing data provenance; (5.49) conceptual model; (5.49) using provenance; (5.38) agent-oriented data curation; (5.38) semantic desktop application; (5.38) modeling provenance; (5.38) browsing provenance log; (4.39) semantic web *Research front*: Balis:eScience07 [14], Balis:PPAM07 [13], Bose:ACMCS05 [36], Bose:SSDBM02 [33], Buneman:FSTTCS2000 [58], Buneman:IDM00 [50], Buneman:SIGMOD07 [62], Cavalcanti:PROV02 [67], Chapman:DBBUL07 [73], Chebotko:escience08 [77], Chen:AHM05 [80], Chiticariu:SIGMOD05 [93], Cuib:ICDE00 [103], Cui:DMDW00 [105], Cui:ICDE00 [104], Cui:thesis01 [102], Cui:TODS00 [107], Cui:VLDB03 [106], DaSilva:DEBULL03 [113], Davidson:DBBUL07 [116], Ding:SS05 [124], Ding:tech05 [123], Fan:AD02 [136], Fan:IOS03 [137], Feng:ICCS07 [139], Fosterb:PROV02 [145], Fox:IJPR05 [152], Frew:PROV02 [158], Frew:SSDBM01 [157], Glavic:BTW07 [178], Greenwood:AHM03 [189], Groth:AHM05 [200], Groth:HPDC05 [198], Groth:IPAW06 [202], Groth:thesis07 [206], Hao:DESI05 [138], Hasan:SSS07 [214], Huang:DEXA05 [223], Hunter:IJDL07 [224], Lawabnia:MSST05 [245], Ledlie:NETDB05 [247], Lord:SWLS04 [255], Macleod:PA2002 [260], Marins:SBC07 [265], Miles:AAMAS07 [285], Miles:AOSE07 [283], Miles:eScience07 [278], Miles:IPAW06 [276], Miles:MASBIOMED2005 [275], Munroe:SEM06 [304], Ram:BO05 [331], Rio:ISVC07 [336], Ruth:ITRUST04 [340], Simmhan:SIGMOD05 [360], Simmhan:thesis07 [358], Spery:GEO01 [368], Stevens:BBIO07 [371], Tan:IPAW06 [377], Townend:AHM05 [385], Townend:ISORC05 [386], Tsai:SOCA07 [387], Vazquez:book07 [390], Wang:ICCS07 [395], Widom:CIDR05 [403], Wong:ISWC05 [404], Zhao:ICSNW04 [416], Zhao:IPAW06 [424], Zhao:ISWC04 [421], ZhaoJ:thesis07 [417], Zhao:SWT03 [415], ZhaoY:thesis07 [422]
Cluster 5: Electronic Logbook 02 (size: 3)
(46.82) multi-scale science; (5.38) supporting emerging practice *Research front*: Myers:SWT03 [306]

Figure 2.7: Labels and Research Front for Clusters 4 and 5

multi-scale science.

2.4 Summary

The use of a co-citation analysis tool has provided us with a new insight on research fronts related to provenance. Research fronts are not structured according to research communities, but take into account the ephemeral and evolving nature of citations, and cross-community citations. While Clusters 4 and 5 consist of early work in the database and workflow communities, Clusters 0 to 3 represent more recent trends. A strong theoretical interest underpins Cluster 1, whereas systems-related issues are the focus of Cluster 2. The explicit presence of the Provenance Challenge inter-operability exercise (in Cluster 3) and its subsequent Open Provenance Model (in Cluster 1) show a growing interest in tracking provenance beyond a single system. The pervasive reference to Semantic Web technologies is also indicative of a growing interest for this type of technology in the community. Finally, while still very small, concerns for security show a growing interest in designing provenance technology that cannot be forged, and

hence can be trusted.

The research fronts that we have identified with the CiteSpace tool have inspired the structure of this survey. To some extent, both the database and workflow communities have similar concerns (though different approaches); thus, we will not structure our discussions on a community-basis. Both approaches have mostly looked at closed systems, but there is growing evidence that they are broadening their investigations to more open environments. Hence, we will look at the possibility of tracking and reasoning over provenance in open environments. In particular, we will look at opportunities and challenges of deploying such a vision over the Web and the Semantic Web. Finally, building on the security approaches, we will look at ways that systems can be made accountable.

Interestingly, the analysis we have undertaken in this chapter is a typical example application that would benefit from provenance. Other scientists would like to be able to reproduce the plots and tables presented in this chapter, they would like to configure the tool differently, or even they may want to apply the same methodology to a totally different set of data. The variety and location of information sources, the number of technologies involved, the kind of processing required (from interactive Java based program, to shell and perl scripts, through manual curation of the database) are a typical driver for the Open Provenance Vision we present in this paper.

Chapter 3

Definition of Provenance

Given that our vision is to track and exploit the provenance of information on the Web, it is necessary to define what we mean by provenance in this context. Luckily, as provenance is studied by multiple sub-disciplines of computer science, various definitions of provenance have been proposed, but unfortunately, several of them are expressed in the context of specific technologies, or under specific assumptions, which do not directly apply to the Web. Instead, with a progressive approach, from dictionary definition to conceptual definition, and by bringing realistic distributed systems assumptions, we aim to propose a definitional framework under which we can develop our vision of provenance on the Web.

Starting with the dictionary definition (Section 3.1), we propose a general and conceptual definition of provenance expressed as a process (Section 3.2). To be more concrete, we adopt a typical Web Mashup application (Section 3.3), which we use to illustrate our conceptual definition, and to review existing alternate definitions, which we cast in the context of the Web (Section 3.4). We then discuss some of the assumptions and considerations typically taken into account in the context of provenance (Section 3.5). Finally, adopting a distributed systems perspective, we make the distinction between the assertion of information related to process, and the extraction of provenance by means of queries (Section 3.6).

3.1 Dictionary Definition

We first introduce the dictionary definition of the word 'provenance'. Its etymology is the French verb 'provenir', which means to come forth, originate. According to the Oxford English Dictionary [1], provenance is defined as follows.

Definition 3.1 (OED Provenance Definition) (i) *the fact of coming from some particular source or quarter; origin, derivation.* (ii) *the history or pedigree*

[1] www.oed.com

of a work of art, manuscript, rare book, etc.; concretely, a record of the ultimate derivation and passage of an item through its various owners. □

Likewise, the Merriam-Webster Online Dictionary [2] defines provenance as follows.

Definition 3.2 (MWO Provenance Definition) (i) *the origin, source;* (ii) *the history of ownership of a valued object or work of art or literature.* □

Both definitions are compatible since they regard provenance as the derivation from a particular source to a specific state of an item. The nature of the derivation, or history, may take different forms, or may emphasise different properties according to interest. For instance, for a piece of art, provenance usually identifies its chain of ownership. Alternatively, the actual state of a painting may be understood better by studying the different restorations it underwent.

From Definitions 3.1 and 3.2, we can also distinguish two different understandings of provenance: first, *as a concept*, it denotes the source or derivation of an object; second, *more concretely*, it is used to refer to a record of such a derivation. We shall return to such a distinction when we define the notion of provenance in the next section.

3.2 Definition of Provenance in Computer Systems

In this section, we focus on data produced by computer systems, published and discovered on the Web; we seek to define the provenance of a piece of data (also referred to as data item or data product). The two dictionary definitions consider provenance to be the derivation from a particular source to a specific state of an item. In computer systems, activities are carried out by processes that take input data, input state, input configuration, and produce output data and output state. Such processes are compositional by nature and can be the result of sophisticated compositions (sequential, parallel, conditional, etc) of simpler processes.

If we have a description of the past process that resulted in a data item, then we are able to explain how such a data item was derived. Hence, following previous work [206, 195], we adopt the following definition of provenance, which makes the notion of process explicit.

Definition 3.3 (Provenance as Process) *The provenance of a piece of data is the process that led to that piece of data.* □

In relation to the two common sense definitions of provenance, we note that Definition 3.3 is concerned with provenance as a concept. From a Computer

[2] http://www.merriam-webster.com/dictionary/provenance

Science perspective, the goal is to conceive a computer-based *representation* of provenance that allows us to perform useful analysis and reasoning to support the drivers for provenance identified in Section 1.1. This representation has to capture details about the process; we will see in Chapter 5, a concrete example of such representation.

As an illustration, the provenance of the plot in Figure 2.1 (page 8) may include the steps involved in producing the plots (the filtering, the data transformation, the plotting program), its parameters, the data involved, the user who initiated the computation, the repositories where some of the data was collected from, their hosting institution, etc.

We should note how broad such a definition of provenance is, since it allows us to capture a variety of activities and data that may have influenced the data item in question. We also note that this definition is totally technology agnostic, and could apply to a plot that may be found in a file system or on the Web, that may have been produced by a workflow or a Java program, that may rely on data found in a public database or in a set of files, or that may have been derived by a single machine or multiple computers distributed across the Internet.

To deal with the complexity of today's applications, which tend to be composed by assembling services and components together, possibly dynamically, systems are studied under specific assumptions. Given these assumptions, it is also not surprising that various definitions of provenance have been proposed in different contexts. To illustrate this and other definitions of provenance, we first introduce an exemplar application.

3.3 Mashup Exemplar Application

We consider a Web 2.0 mashup application that includes a map displaying information extracted from feeds from multiple sources (e.g. RSS feeds, Twitter, Facebook). The user interface allows for information to be filtered according to multiple criteria: user trails, time intervals and geographical regions. The application caches (some) data locally. As the user makes a selection, information is obtained from the different caches according to the selection and mashed up on the map.

We can see a variety of technologies exploited in this application, whose architecture is illustrated in Figure 3.1. The mashup service renders Web pages, whose contents are computed by a query engine (Q), operating over local caches, populated by RSS feeds from several information providers (Trails provider, Photos provider, and Blogs provider). Each provider may use their own technology: for instance, the blog contents may consists of XML documents, whereas the photo repository consists of files, stored in a directory and exposed to the Web.

After the user selects `alice` (one of the three individuals tracked by the Trails and Blogs providers), the mashup displays a Southampton map with a picture of

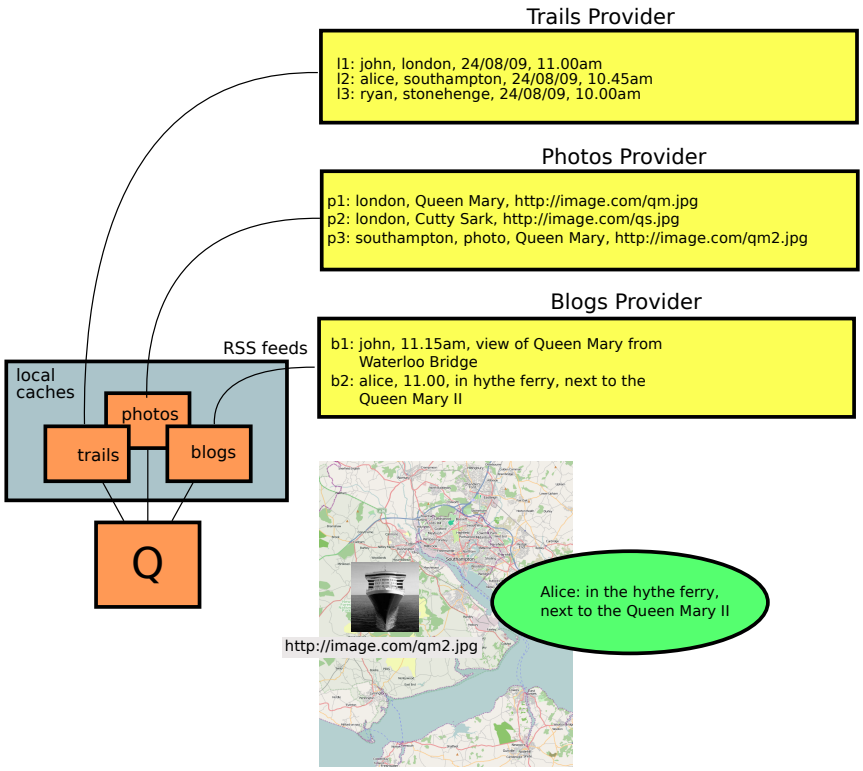


Figure 3.1: Provenance in a Mashup

the cruise liner Queen Mary II in Southampton harbour, with the blog comment "in hythe ferry, next to the Queen Mary II". Initially, the various information items displayed in the mashup are obtained by running the following query over the local cache:

```
select photos.url,  blogs.comment
where trails.user='alice'
 and  trails.location=photos.location
 and  blogs.user=trails.user
```

The query outlined above is to be seen as a "continuous" query running over the local cache, so that whenever more data is pushed by information providers, the mashup is automatically refreshed.

3.4 Alternative Definitions of Provenance

We now illustrate various definitions of provenance using this example application.

Provenance as Process Following Definition 3.3, the provenance of the mashup (whose content is depicted in Figure 3.1) is the process that resulted in this content being displayed. The final mashup was produced by combining a map with the result produced by the query engine, itself running a query based on a user selection and extracting data from the Photos and Blogs cache. According to this definition, any data, event or user action that can be connected to the mashup through a computational process potentially belongs to the provenance the mashup.

Provenance as a Directed Acyclic Graph There is a recognition[3] that provenance can be expressed by a directed acyclic graph (DAG), explaining how a data product or event came to be produced in an execution [298]. In a first approximation, we can consider that in such a DAG, nodes represent data items and edges data derivations. An illustration of provenance as a directed acyclic graph is displayed in Figure 3.2 for the mashup application. The mashup instance appears at the bottom. Each node represents a data item that existed at some point in the computation. For instance, `b2` denotes the eponymous tuple in the Blogs provider, while `ba2` represents its copy in the local cache, and `bb2` represents the copy extracted by the query engine; the query itself is represented by node `q`, which causes copies of `p3` and `b2` to be incorporated in the mashup. In this example, by design, the granularity, i.e. smallest information we track provenance for, is the tuple; alternatively, we could have decided to adopt table or cell as the granularity.

Why-Provenance Initially defined in the context of databases, data lineage (Cui and Widom [107]) and why-provenance [59] identify the set of tuples, whose presence justifies a query result. In our exemplar application, the image and the blog comment displayed in the mashup are justified by the information records `t2`, `p3` and `b2`, from the Trails, Photos and Blogs providers respectively. In Figure 3.2, why-provenance is represented by computing the transitive closure of plain edges, representing data dependencies.

Where-Provenance The user may realise that a spelling mistake occurs in the bubble displayed by the application (the word 'hythe' is not capitalised). To correct it, the user would like to know the field, record, and database in which this string originally occurred, so that it can be updated accordingly. In our case, this is in information record `b2`, in the Blogs system. Such a notion of provenance, referred to as where-provenance [59, 55], was initially defined in the context of databases; it helps identify where information was copied from. In Figure 3.2, where-provenance is obtained by following `sameAs` edges, which represent data being copied or shared.

[3] http://twiki.ipaw.info/bin/view/Challenge/SecondWorkshopMinutes

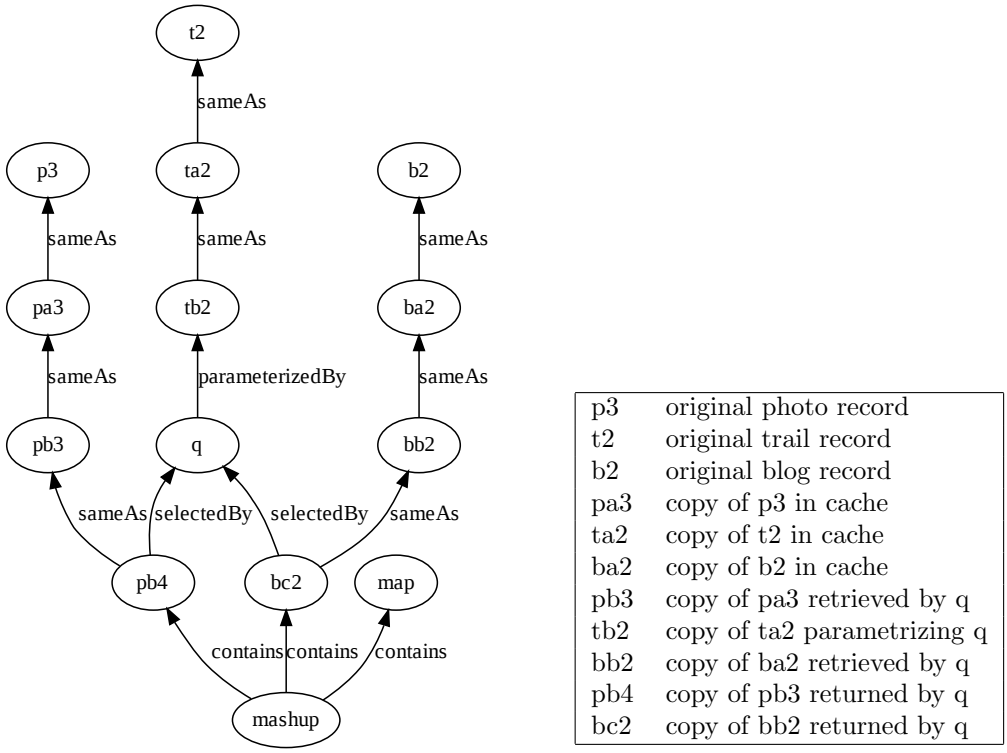


Figure 3.2: A Data Derivation DAG for the Mashup

How-Provenance Why provenance tells us which source tuples witness the existence of a result, but they do not tell us how they are involved in the creation of this result, i.e. how their involvement proves the result. How-provenance [188] consists of a polynomial representation that hints at the structure of the proof explaining how an output tuple is derived. How provenance was defined in the context of relational algebra and recursive datalog. A very approximate representation of how provenance to our exemplar application is the polynomial $\texttt{p3} \times \texttt{b2} \times \texttt{t2}^2$, which explains that `p3`, `b2` and `t2` are all required to justify the presence of `pb4` and `bc2` in the mashup, but in addition `t2` was used twice, to select both `p3` and `b2`. How-provenance can also be derived from the process representation of Figure 3.2, since such a kind of polynomial can be composed from the graph representation by considering tuples as variables. Given that, in the most general case, the representation of Figure 3.2 is a directed acyclic graph, which would result in exponents (greater than one) in the polynomial whenever several distinct paths lead to a same tuple.

Glavic [176] introduces an alternative terminology for these notions: *copy-*

contribution for where provenance, *input-contribution* for why-provenance, and *influence-contribution* for Cui and Widom's lineage.

Provenance as Annotations There exist ontologies, such as Dublin Core [449], to provide structure and semantics to metadata of resources. Aspects of these ontologies are provenance related, such as author, creation date, and version. Such information can also be seen as a specialisation of Definition 3.3, as it is concerned with specific properties of past processes. Miles is proposing a mapping[4] of the Dublin Core to the Open Provenance Model.

Other definitions Other definitions are encountered in the literature, but they are not as popular, in part, because they do no explain how a given data product has been derived. For example, Hasan *et al.* [215] adopt an event-oriented view, according to which a provenance chain for document D is defined as a non-empty time-ordered sequence of provenance records, where a provenance record captures an access to a document D.

3.5 Assumptions

Given the broadness of Definition 3.3 and the universal appeal of provenance, work has independently been undertaken in multiple communities, using different assumptions. We review some of them here.

System Scope Some approaches have a working hypothesis that a given system entirely manages the flow of information, and that provenance has to be tracked within the scope of that system. Examples of such systems include operating system [218][159], desktop [265], statistical packages [143, 21], databases [59, 406, 176], data warehouses [104], and workflow systems (Kepler [39, 41], Taverna [419], VisTrails [156], VDS [424], Pegasus [278]). Other approaches allow for varying degrees of open-ness: curated databases allow for user-edits [52], PASOA allows for service-oriented architectures [195] where the number of components can dynamically change and is not known at design-time, or some allow for data to be published on the Web, with data and provenance both accessible by simple browser navigation [447]. Others even aim at tracking the provenance of objects in the physical world [231].

Program Some approaches assume that both the programming language and the program that executed the process are known, and therefore can be used

[4] http://twiki.ipaw.info/bin/view/OPM/ChangeProposalDublinCoreMapping

to identify provenance [59], to derive a reverse function that computes provenance [406], or to encode provenance efficiently [424][419]. In general, the benefits of such approaches is that provenance can be precisely defined in terms of the execution semantics, and that it can be efficiently inferred or represented. The downside is that it may require the reasoner that wishes to process this kind of provenance to have an understanding of the language used for this execution. As a result, this type of reasoning may not be portable across environments (for instance, a reasoner operating over VDS provenance will not be able to reason over Taverna's). Other approaches do not make such an assumption, but therefore rely on an ontological description of what happened, e.g., OPM [299] and PASOA [195]. In that case, the reasoner needs to access the ontology definition to reason over provenance.

Granularity In both business and science, applications have to manipulate collections of data, e.g., structured data, sets, hierarchies, tables, rows, nested collections, files, or directories. Approaches to tracking the provenance vary according to their ability to track the provenance of collections and their members. Relational databases have usually tracked provenance to the level of cells, whereas in some operating systems and workflow systems, the trend is to deal with files (e.g., PASS [218], VDS [424], ES3 [161, 159]), but recent approaches are dealing with collections and their members explicitly (e.g., OPM [292, 299], Kepler [39, 41], Taverna [419]).

What is in the provenance? The broad definition of provenance (Definition 3.3) allows for a vast range of information to be captured, including data derivation, libraries, hardware where the computation was run, and runtime information. Others are most prescriptive, and identify data items that are the original raw value as it entered a system [55], data items that were the cause of a given data [107, 176, 59], and variants where a summary of how the data were used is incorporated [188].

The provenance of what? All the definitions above are concerned with the provenance of data products. Some authors generalize this notion to computational processes. For instance, Michlmayr *et al.* [274] propose the concept of service provenance. By this, they mean information such as past quality of service and past invocations as provided by a service monitor. We believe that such service provenance can be framed in terms of Definition 3.3, by considering the provenance of the service state, given by all its previous interactions and initial configuration. Alternatively, Freire *et al.* [156] introduce the concept of workflow provenance (discussed in Section 4.5); workflows are a specific kind of data for which the process of derivation also needs to be tracked. Hence, for the rest of this survey, we will only consider the provenance of data.

Time It is generally recognized that a provenance model does not have to include time [292]. Indeed, considering the graph of Figure 3.2, data derivations are represented by edges explicitly. If creation time of data items was known, there would be an expectation that if a data item is derived from another, the former would be created after the latter. However, the converse does not hold: time precedence does not imply data derivation. While time is not mandatory, it is perceived that it is practical for users to be able to refer to time [298]. Therefore, most systems support a notion of time, so that users can refer to executions or data products according to the time at which they took place or were produced.

3.6 Provenance: a Query over Process Assertions

Let us again consider the mashup application of Figure 3.1. An observer could observe all activities in detail, and produce the mashup's provenance as displayed in Figure 3.2. This observer would have to be distributed (since it has to inspect all activities taking places at the different sites across the Web involved in this computation) and it would have to be trusted to be able to inspect all systems. In fact, it would have to be omniscient to be able to observe every detail of the computation. While such an omniscient and ubiquitous observer could be conceived in specific cases, such as in a monolithic application, it is simply not possible for such an observer to exist across the Web.

Since it is impossible to forcibly observe the behaviour of arbitrary systems, instead, one would have to rely on *assertions* made by the systems' distributed components, about their local actions and involvement in a computation. The description of a local computation by one component would have to be connected with other descriptions of local computations by other components according to the distributed flow of information, in order to formulate an overall meaningful description of a distributed computation.

Therefore, we regard the graph displayed in Figure 3.2 as the result of a query over a set of assertions made by the different applications about their involvement in the computation; the query reconciles and composes assertions according to the flow of information. The PASOA approach [206, 195] makes the distinction between assertions about a process (referred to as p-assertions) and provenance obtained by a query over process assertions.

This conception of provenance as the result of a query is further justified, by the following. The mashup's provenance consists of a graph that traces back to the tuples `t2`, `p3`, and `b2`. These information items also have a provenance: the trail information could be derived from a GPS log, the photo provider could be a curated database. There is no natural reason to terminate the provenance graph

at `t2`, `p3`, and `b2`. If we were to study the cache content, it would be perfectly valid to terminate the graph with tuples stored in the cache. If we care about raw data capture, we would have to trace back to the various sensors involved in producing the data (GPS and camera). In fact, provenance needs to be scoped according to the user's interest; otherwise, by default, the provenance of any item would conceptually trace back to the Big Bang, marking the origin of our Universe. The scope can identify the systems in which the tracing back should terminate, or the type of source data we are interested in identifying. Miles [276] proposes such a scoping mechanism to allow queriers to delineate the provenance of data items.

Interestingly, with this novel understanding of provenance, we can now revisit the alternate definitions of provenance of Section 3.4. These definitions can be seen as predefined queries over process assertions; for instance, where-provenance as a query that always follows the "sameAs" edges in a provenance representation.

3.7 Summary

By adopting a general definition of provenance as a process, we have converged towards a conceptual definitional framework for provenance that can apply to information flow over the Web. Concretely, provenance needs to be represented explictly in a computer-processable format to allow for reasoning, so as to support the drivers discussed in Section 1.1. The proposed definition can be seen as a generalisation of alternative definitions encountered in the literature.

We have also made the distinction between assertions about processes and provenance resulting from a user-specific scoped query over such assertions, with respect to a specific data product. This general view of provenance is instantiated and optimised in different systems. Assertions can be optimised for specific execution environments, and queries over assertions can be pre-defined in systems; furthermore, queries can be eagerly or lazily computed, and query results can even be published as "provenance-metadata". The distinction between process assertions and queries also allows us to appreciate the lifecycle associated with provenance: process assertions need to be collected and accumulated as computations proceed, possibly without knowing which data product is ultimately to be derived. Once accumulated, these assertions can be queried to provide novel functionality to users.

Chapter 4

Provenance in Workflows and Databases

The majority of work on provenance has been undertaken by the database, workflow and e-science communities. Since several good surveys already exist on provenance in databases [226, 380, 379, 178] and in workflows and e-science [360, 297, 36, 116, 112], we do not intend to repeat these here. Instead, our purpose is to contrast the work between these two communities, by identifying similarities and differences.

To understand the distinct approaches by these communities, it is crucial to appreciate how differently workflow and database technologies are exploited. We begin with the scientific context, and then consider business. Workflow technology is increasingly considered a rapid experiment development tool, with workflow modifications, frequent runs, and parameter tuning [174]; workflow languages are a mechanism to rapidly glue libraries and services, easily transform data, and rapidly automate computational activities. Hence, a primary driver for provenance is reproducibility of scientific analyses and processes. Furthemore, provenance is not only used for interpreting data and providing reproducible results, but also for troubleshooting and optimization [278]. For an extensive analysis of user requirements for provenance in e-Science, we refer the reader to [279]. We note that the use of workflows in business differs substantially. Business workflows are less of an iterative development tool, but are used to implement business processes; in such a context, traceability and accountability are important concerns [108].

In the scientific context, databases have traditionally been used for archiving data [49]. Some of these databases undergo frequent updates, and new versions are released regularly. In particular, curated databases are constructed by the "sweat of the brow" of scientists who manually assimilate information from several sources [51, 10]; they may be the result of a great deal of manual annotation, correction and transfer of data from multiple sources. In that context, provenance information concerning the creation, attribution, or version history of such

data is crucial for assessing its integrity and scientific value[1]. A series of technical requirements, applying to both workflow and database technologies can be found in [73].

In the most general cases, workflows are used to compose services whose detailed behaviour is not necessarily known. Given that provenance in such a context involves components regarded as black boxes [98], Tan [380] categorizes it as workflow and coarse-grained. We note however that it need not be the case. More or less detail of such services can be exposed, making them grey boxes [116]. Thus, we believe that there is not two competing solutions to provenance, fine-grained vs coarse-grained, but there is a continuum of solutions where all parts of the systems in which information flows can be exposed in a fine or coarse manner.

Figures 4.1 and 4.2 display tag clouds for database and workflow publications from the provenance bibliography. They share a common focus on scientific activities, workflows, provenance queries, and building and evaluating systems.

Figure 4.1: Database Cluster Tag Cloud

The rest of this chapter compares and contrasts approaches, from various perspectives, which are crucial to establishing the foundations of provenance on the Web. *(i)* For users to deal with the amount of information contained in provenance, mechanisms are required to abstract and synthesize information in views customized to users (Section 4.1). *(ii)* A specific aspect of abstraction is concerned with collections of data (Section 4.2). *(iii)* If the provenance of everything is to be tracked, consideration should be given to storage requirements (Section 4.3). *(iv)* The means to actually query provenance need to be provided (Section 4.4). *(v)* Tracking the evolution of workflows is a special kind of

[1]This kind of requirement is evidenced by journals, such as `http://www.earth-system-science-data.net/`, which adopt metadata schemas to annotate data in order to ensure their quality and potential reuse.

Figure 4.2: Workflow Cluster Tag Cloud

provenance tracking (Section 4.5). *(vi)* Formal properties of provenance are now emerging (Section 4.6). *(vii)* Finally, many activities involve humans in the loop, who impact on decisions and processes, and therefore need to be made explicit in provenance representations (Section 4.7).

4.1 Views and Abstraction

The provenance of a data product may be large, in particular, when the data product is the result of a long and complex computation. This presents challenges to users since it becomes very difficult for them to understand such provenance and make sense of it. Hence, novel techniques have been devised to allow users to deal with the complexity and size of provenance information. Some of these mechanisms consist of groupings of data, which we discuss in Section 4.2. In this section, we focus on approaches that have been proposed to structure historical information. They consist of layering, workflow-induced views, tracers, and accounts.

Redux [17] offers a four-layer provenance model. The first level consists of an abstract description of the experiment that captures abstract activities in the workflows and relations among them. The second level represents an instance of the abstract model, which captures instances of activities and additional relationships, as classes of activities are instantiated. The third level captures information to trace the execution of the workflow, including input data, parameters supplied at runtime, branches taken, and activities inserted or skipped during execution. The final level represents runtime-specific information, such as the start and end time of workflow execution, start and end time of individual activity execution, status codes and intermediate results, information about the internal state of each activity, along with information about the machines where activities were

allocated.

A similar layering is adopted by VisTrails [348]: the workflow evolution (to be further discussed in Section 4.5) layer captures the relationships among the series of workflows edited by the user; the workflow layer consists of specifications of individual workflows; and the execution layer stores run-time information about the execution of workflow modules (e.g., execution time, machine name, date, etc.). The Wings/Pegasus workflow system [232] introduces the notion of reusable workflow template that is instantiated into a workflow instance, containing execution details. Provenance is structured in a similar manner, and, in essence, follows a layering that resembles that of provenance in Redux and VisTrails.

ZOOM [99, 98, 30] builds on the concept of composite step-classes - or sub-workflows - which is present in many scientific workflow systems to develop a notion of *user views*. There are several reasons why composite step-classes are useful in workflows. First, they can be used to hide complexity and allow users to focus on a higher level of abstraction. Second, composite step-classes can represent authorization levels; users without the appropriate clearance level would not be allowed to see the details of a composite step-class. A partial ordering on user views can be defined using the containment of step classes. Such views can be referred to when querying provenance: a user view determines what level of sub-workflow the user can see, and thus what data and tasks are visible in provenance queries [30]. The challenge is to construct such user views dynamically: a bottom-up approach to constructing well-formed user views is described in [29, 30].

Both VDL [97] and Karma [363] rely on a notion of nested workflows, which allows provenance to be grouped, and retrieved according to its depth. We discuss in Section 6.1, the technique that Hunter and Cheung [91, 224] propose to create user views by relying on Semantic Web techniques.

Tracers [282, 201] are unique tokens propagated by services at execution time through interactions, very similar to transactional contexts passed in distributed transaction systems. Tracers are typed and can be given different semantics: tracers can be used for instance to delimit a workflow run or to capture the dynamic nesting of workflow execution. As tracers are communicated by the application, they are also documented among the assertions made about execution. By this mechanism, a tracer can be used to delimit a (sub-)process or activities with some properties, and so can help bundle all process documentation regarding that (sub-)process or activities. Provenance query interfaces also take tracers as input to ensure that provenance information that belongs to the scope of a tracer is returned. Tracer-based views need not be hierarchical (as opposed to the workflow-induced views); for instance, they could be location-based.

The Open Provenance Model [299] (whose details are provided in Chapter 5) offers the notion of an account. Accounts are a workflow-independent mechanism to introduce abstraction and structure in a provenance trace. Accounts allow for multiple descriptions of a given execution to co-exist in a provenance trace. Such

accounts can overlap (meaning they are related to a same execution), they can be hierarchical and linked by a notion of refinement[2], or non-hierarchical. In the latter case, they may even be offering conflicting views about a same execution observed by two different observers. The interest of accounts is that they are independent of the technology used to run the application, and therefore apply to workflow and non-workflow based systems.

4.2 Data Collections and Streams

Users very frequently have to deal with collections of data, as opposed to individual data items. Such collections may be more or less structured; they can be sets (e.g., file directories), relations (e.g., SQL relational tables), hierarchies (e.g., XML documents), or arrays (e.g., two-dimensional matrices of numbers). A collection of data is a group of data items, generally of the same type, which may be ordered or non-ordered. The grouping tends to reflect some properties of its elements: all the results produced by an experiment, all the results returned by a query, all the photos taken by a given camera during a period of time, all the simulations results regarding a given project, or a set of URIs returned by a search. For end users, collections become first-class entities that can be annotated, manipulated, transformed, or archived. As far as provenance is concerned, it is therefore important to distinguish the provenance of a collection from the provenance of its individual members. The provenance of a collection constitutes a form of abstraction, similar to the ones discussed in Section 4.1, where the collection's provenance abstracts away from the details of its members' provenance. Representing the provenance of collections and their members is challenging, when we consider all the potential dimensions of the problem: collection mutability, granularity and efficient representation. We now discuss several approaches tackling these problems.

Many workflow systems tend to create new data, without ever updating existing ones; this also goes for collections, where such systems, process input collections, mapping operations on their individual elements, and producing new collections. Examples of such systems include VDL [97] capable of manipulating directories of files, or Taverna/myGrid offering mapping functions operating over collections. Kepler [7] also allows collections to be manipulated, but such collections are stateful since Kepler provides operations to delete and insert members of a collection.

Database technology naturally deals with collections: for instance, relational tables, rows and cells are the constituents of the relational model, whereas hierarchies are at the core of XML databases. Provenance in databases has been dealing with data collections at various levels of data granularity: the seminal paper on why- and where- provenance [59] deals with semi-structured data (XML

[2] http://twiki.ipaw.info/bin/view/OPM/ChangeProposalMultipleHierarchicalRefinement

databases), whereas [55] deals with the provenance of cell contents in the presence of updates, and [107, 176] deal with tuples in SQL databases. Generally, these approaches allow for provenance to be tracked at all levels of data granularity: for instance, in [55], annotations to tables, rows and cells are propagated.

A challenge for systems constructing an expressive representation of provenance is the size of the provenance of a collection and its members. In the context of the Open Provenance Model, the collection profile[3] allows for the provenance of members to be derived from the provenance of collections, by applying inference rules that are specific to the operation performed on the collection, such as map or filter. Instead, Anand *et al.* [7] prefer recording changes performed to data structures; a motivation for their choice of representation is storage efficiency, which we further discuss in Section 4.3.

So far, this section has discussed the concept of collection seen as a first-class aggregation of data, which may vary over time as elements are added or removed, but is persistent: at any point in time, it is possible to retrieve a collection and obtain all its members. On the other hand, streams are collections that are typically ephemeral (i.e., not made persistent or archived) and temporal. We discuss now specific techniques for handling the provenance of streams.

Sensors and data streaming techniques are increasingly used in a wide range of applications, from science (weather forecast [393]) to health care (remote health monitoring [287]). Such streamed data are used by sophisticated simulation, modelling and analysis tools. Streamed data which are of interest to both workflow and database communities, carry specific problems related to provenance. Examples of provenance queries are: which sensor was a piece of data produced by? What transformations were involved in deriving a stream? Which events were ancestors of a given event in a stream?

Vijayakumar and Plale [393, 392] adopt a stream as the unit of data they track provenance for, so as to ensure a lightweight and low overhead approach to provenance recording in distributed stream applications. The downside is that the granularity of their provenance model is so coarse that it is unable to answer queries related to individual stream elements.

The goal of the Kepler workflow system is to build streaming applications; the provenance model conceived by Anand *et al.* [7] consists of changes performed to data structures; it is itself a stream, which is embedded in application streams.

Misra *et al.*'s approach [287] is fine-grained since, given an output (e.g., a medical alert) generated by a stream processing application, their system not only recreates the processing graph that generated the output, but also provides all the elements of the intermediate data streams that generated it. To this end, they introduce the Time-Value-Centric (TVC) model [287, 396, 228] which is an algebraic approach to compute all the ancestor events an event depends on.

[3] http://mailman.ecs.soton.ac.uk/pipermail/provenance-challenge-ipaw-info/2009-June/000120.html

While this section focused on streamed applications typically produced by sensors, work has also been undertaken on provenance in the context of video streaming applications. Gehani [170] designed an algorithm for in-band encoding of lineage metadata in video streams.

4.3 Efficient Storage of Provenance

Provenance can become huge: in the public database Gene Ontology, the provenance of a single tuple has been observed to be 10Mb [332]; likewise, a 250Mb database of biological data is associated with 6Gb of provenance [73]. We should note from the start that there is a trade-off between compact representation (reducing recording/upload time), compact storage (reducing storage requirements) and query time.

Barga and Digiampietri [17] observe that a layered model can expose opportunities to store provenance traces efficiently in a storage manager. This point is also noted by Scheidegger *et al.* [348], since structuring the provenance information into layers leads to a normalized representation that avoids the storage of redundant information.

Anand *et al.* [7] observe that in typical workflow steps, it is the case that not all outputs directly depend on every input. Therefore, to accurately trace the provenance of data, fine-grained descriptions of data dependencies need to be asserted. In the presence of collections and nested collections, the size of such descriptions can be considerable, resulting in poor recording performance and high storage requirements. They therefore propose a compact representation of provenance, which essentially tracks the changes performed to data structures, assuming that any other element remains identical. This compact representation is associated with inference rules, allowing all dependencies to be explicitly derived. Such representation techniques are combined with a series of storage representation optimizations. Through a range of real and synthetic benchmarks they compare the efficiency of their representation techniques, from recording, storage and querying perspectives.

Observing the increasing storage requirements for provenance, Chapman *et al.* [75, 72] identify two families of techniques to decrease the storage needs for provenance: factorization processes and inheritance-based. Factorization techniques factor out common "sub-expressions" in the provenance of different items, allowing them to be stored once for each item. Alternatively, an orthogonal optimization is based on similarities in a local portion of data collections (Structural Inheritance) or between the provenance associated with data items of a particular type (Predicate Inheritance). When provenance can be inherited by an item, there is no need to record any provenance with that item; instead, the inheritance mechanism can correctly instantiate what is required. Such techniques can reduce storage requirement by a factor of 20, while provenance remains queryable.

Ré *et al.* [332] note that is is often unnecessary for systems to track all derivations. Indeed, sometimes, an approximation returned quickly is more valuable than the accurate complete provenance returned after a long time. Furthermore, complete provenance does not identify the most influential steps in that derivation. As a result, they propose *approximate lineage* as an alternative to complete provenance. It compresses the provenance by tracking only the most influential facts in the derivation. They introduce two forms of approximate lineage. In their database approach, complete lineage is represented as a boolean formula over a set of boolean variables. Sufficient lineage is a smaller formula that logically implies the original. The second approach is polynomial lineage, consisting of a real-valued polynomial over boolean variables. (We note that some form of polynomials is also used in how-provenance [188].)

Buneman *et al.* [49] discuss the overhead of maintaining multiple versions of a data record in scientific databases, in the presence and absence of compression. More recently, a similar overhead study is undertaken in the context of curated databases, where Buneman *et al.* [51] investigate the storage requirement and associated overhead for their copy-and-paste provenance model.

Groth *et al.* [194] take advantage of the properties of a specific class of scientific workflows to derive an optimized provenance representation.

4.4 Querying Provenance

In Section 3.6, we have discussed the idea that provenance is the result of a query over a set of assertions about execution. This approach is adopted by many systems (Taverna [419], PASS [218], ES3 [161, 159], PASOA [276], VDL [97], Kepler [39, 41]) which accumulate information about processes during their execution, and offer query interfaces to retrieve provenance. The first Provenance Challenge [298] shows that a wide variety of standard querying technologies are used: SQL, XQuery [432], Xpath [436], SPARQL [450]. The downside is that implementers expose their implementation schema to their users, which makes it difficult to change it in the future. As an alternative, a series of domain specific languages are being designed to retrieve provenance (and are generally implemented as a translation to standard query languages). These domain-specific languages aim to improve expressiveness by offering new constructs and new abstractions to facilitate the writing of complex queries. We review some of the characteristics of these provenance-oriented query languages.

The PASOA query interface for provenance [276, 195] has two constituents. First, it requires the querier to identify the data item they want to retrieve the provenance of. Second, it mandates a specification of the part of the process the querier is interested in obtaining a description of.

As far as data products are concerned, there are generally two approaches. Several systems, typically integrating workflow execution and provenance, name

all intermediary results with a unique identifier, which can then be used to obtain their provenance. For instance, Taverna/myGrid uses Life Science IDentifiers (LSIDs) [418], Swift uses tag URIs [97], and VDL [424] uses filenames. We refer to this kind of identification as *extensional*, since data items are explicitly enumerated and named when issuing a provenance query. Alternatively, PASOA [276] identifies objects *intensionally*, with respect to workflow steps: for instance, the object contained in a collection passed as input to a workflow step, carried out by a specific service. The PASOA provenance query interface allows for such extensional descriptions to be specified as an XPath expression, over the set of process assertions.

Having identified a data item (intensionally or extensionally), one then needs to select the part of the process we are interested in. We previously indicated that provenance takes usually the shape of a directed acyclic graph; therefore, a provenance query involves some form of transitive closure. Holland *et al.* [217] review the suitability of query languages to address this kind of queries. The PASOA approach [276] offers various ways of specifying process scope: it can be delimited by location (of activities or provenance stores), by types of derivation, or by the type of intermediary data involved in the computation.

The database community has identified several forms of provenance (why, where, how, lineage). When these notions are transposed to a broader context beyond databases, as in the mashup of Chapter 3, a scope also becomes useful. For instance, in Figure 3.2, we may want to identify *where* a tuple originally appears in the cache.

Anand *et al.* [8] propose a Query Language for Provenance (QLP) designed to be independent of any particular physical representation, and that includes constructs tailored specifically for querying scientific workflow provenance. QLP consists of two components: lineage queries (aimed at traversing the transitive closure of the graph), and structural queries (aimed at traversing nested collections); both can be combined into so-called hybrid queries.

The VisTrails provenance query language [348] (vtPQL) is designed to take advantage of the structure of the VisTrails layered provenance. Each level of the query is a simple SQL-like expression with some additional functions, predicates, and attributes. Basic operations that are useful for common querying tasks over workflows, and that further simplify the query syntax have been identified.

Heinis and Alonso [216] show that workflows with a tree structure produce lineage dependencies that can be efficiently stored and queried using interval encoding. They define a provenance query as the transitive closure over a DAG. By a series of benchmarks, they show that recursive queries require little space but can be slow, whereas storing all paths leads to faster queries but increases storage requirement significantly. By using intervals to represent trees, provenance of a node can be determined by finding all the intervals that enclose the interval of this node. They explore how arbitrary DAGs can be transformed into equivalent DAGs that can be encoded with one-dimensional intervals.

Buneman *et al.* [55] take a formal view of expressiveness. They propose an implicit provenance semantics where each data item is enriched with an annotation (a "colour" that represents where it originally appears for the first time in the database). They make an interesting observation: some queries that are regarded as equivalent under traditional semantics, and therefore return identical views, become distinct under implicit provenance semantics because, despite returning data items that have the same value, these items have a different provenance. This definitely raises issues for query planners and optimisers, since they should be provenance preserving.

In contrast to why-, where-, and how- provenance, Chapman [72] investigates a *why not* query[4]. This query produces a series of statements about the potential reasons the data of interest to the user is missing from a result set. Two algorithms answering this query by traversing the provenance graph in a forward and backward manner respectively are proposed and compared in evaluations.

Bao *et al.* [15] introduce an algorithm for differencing provenance (due to workflow execution). The difference or edit distance between a pair of valid runs of the same specification is defined as a minimum cost sequence of edit operations that transform one run to the other. While the differencing problem is NP-hard for general graphs, a polynomial solution is proposed for series-parallel graphs (with nested loops), capturing a broad class of scientific and business workflows.

4.5 Workflow Evolution

In the introduction of this chapter, we discuss how workflow technology is increasingly used for extreme experiment design by e-Scientists. Given that frequent tweaks to workflows and modifications of parameters can have a significant impact on experimental results, they necessarily need to be included in the provenance of the results.

Thus, tracing the provenance of workflows has been an increasing concern. Since the workflow and its parameters are themselves a data set, Definition 3.3 still applies: the aim of this research is to track the process by which a workflow has been derived. This is the approach taken by VisTrails [156, 252] which maintains workflow provenance, by capturing modifications made to workflows through its integrated development environment. This allows scientists to easily navigate through the space of workflows and parameter settings used in a given scientific task. In particular, this gives them the ability to return to previous versions of a workflow and compare their results.

In some cases, the workflow does not evolve because of user modifications, but because of worklow compilers. The Pegasus workflow compiler [278] takes as input an abstract workflow (specified as a DAG) that is compiled into an instantiated

[4]This form of query in the context of provenance was first brought to the author's attention by Yolanda Gil in 2006.

workflow, directly executable by a workflow engine, where computation location, data transfers and libraries to invoke have all been made explicit. In this case, provenance can be used for troubleshooting and understanding runtime behaviour of the workflow system and of the application; this is achieved by connecting runtime information to the original abstract specification, as designed by the user.

4.6 Provenance Semantics

A large proportion of the work on provenance in the database community has been of a theoretical nature. Some of these results have applications beyond the database world and should be considered as desirable approaches to provenance semantics in general. Buneman *et al.*'s *implicit provenance* approach [55] considers the semantics of a query language (nested relational calculus) where values have been tagged by an annotation, referred to as a colour, denoting the origin of that value. As values are propagated, the language passes along annotations. As a result, for any result produced by a program, the associated annotations indicate where the value was derived from.

A similar approach is also adopted by Souilah *et al.* [366], who introduce a formalism for provenance in distributed systems based on the π-calculus. Its main feature is that all data products are annotated with metadata representing their provenance. Here, annotations consist of sequences of send- and receive-events, that are extended whenever values are communicated by the application. Souilah *et al.*'s annotations are richer than Buneman *et al.*'s colouring scheme, which means that more sophisticated provenance queries can be answered. Souilah *et al.* define a notion of correctness of a provenance annotation if what it tells us about the past of a value agrees with what actually took place.

Cheney *et al.* [88] introduce a notion of provenance trace, and investigate some of its properties. Two of them are worth noting: a trace is *consistent* if it describes what happened during execution. It has the *fidelity* property if it contain enough information to describe how the program would have executed with different inputs.

Moreau *et al.* [296] formalise the Open Provenance Model [299] and investigate its expressivity. They formally specify the kind of inferences the model allows.

Glavic and Alonso [177] demonstrate that the widely used definition of why-provenance (as defined by Cui and Widom [107]) fails in the presence of nested subqueries. They show that in the presence of a subquery, provenance includes tuples that do not actually contribute to the result (false positives), and furthermore that there is ambiguity in the presence of multiple subqueries. They propose a revised definition of provenance that tackes this limitation. Their solution is implemented in the Perm system, which rewrites SQL queries to propagate provenance alongside query results (similarly to [55]). For a given query [176],

Perm generates a single query that produces the same result as the original query but is extended with additional attributes used to store provenance data. The benefit is that the rewritten query is expressed in SQL and can be optimised by the DBMS, and the provenance information is represented in relational tables, not requiring the data model to be extended.

4.7 Human-Driven Workflows

There are a number of domains, where the workflow is not driven by an automatic workflow enactment engine, but directly by humans. Provenance in that context is also important. We review the benefit of provenance in forensic analysis, knowledge discovery and visualization.

Levine and Liberatore [250] seek to improve the reproducibility and comparison of digital forensic evidence. They propose a simple canonical description of digital evidence provenance that explicitly states the set of tools and transformations that led from acquired raw data to the resulting product. This provenance representation allows for the comparison and the reproduction of results.

The knowledge discovery process is indeed a process and the steps that a user takes to discover knowledge are as important as the knowledge itself [192]. Groth *et al.* [192] propose to recognize the user interactions and annotations as first-class objects, which can be exploited by other users to discover resources. Information captured in the visualization system allows for visualizations to be replayed and for previous discoveries to be found again.

Silva *et al.* [356, 20, 65] use VisTrails' action-based provenance model to capture changes to parameters and pipeline definitions to ensure that users are able to reproduce visualizations, and to let them easily navigate through the space of pipelines created for a given exploration. Likewise, Jankun-Kelly [227] derives a visualization process graph representing visualization activities. A classification of these graphs is introduced and metrics to analyse them are defined.

Gotz and Zhou [186] introduce the notion of *insight provenance* to refer to a historical record of the process and rationale by which an insight is derived during a visual analytic task. Instead of relying on an action-based provenance model as in VisTrails, they aim to capture automatically a semantic record of user activity, consisting of high-level descriptions of analytic actions. Such actions are themselves inferred from low-level user interactions.

4.8 Summary

The bulk of the work on provenance has been undertaken by the database and workflow communities, specifically in the context of scientific applications. While database and workflow technologies are used significantly differently by scientists,

they share numerous provenance concerns. As important progress is being made on the theoretical front, with various semantics of provenance, there are still a number of challenging issues to consider. We have reviewed the problems of creating user-adapted views that abstract away from tedious details of execution or of manipulated data sets, expressive provenance queries, and efficient provenance storage. Furthermore, processes can be driven by humans, and provenance should reflect this in its representation.

Chapter 5

The Open Provenance Vision

In the past, applications used to be monolithic, running within a single security domain, possibly a single machine, without having to inter-operate with any other software system. Today's applications are substantially different: they consist of many components, typically involving multiple technologies, deployed in separate security domains, and architected according to service-oriented principles [448], promoting loose coupling and reuse. Furthermore, with the advent of cloud-computing [433], many applications are architected around the Web, publishing and discovering information over the Web, mashing it up, and republishing it. In this context, the challenge is to be able to track the provenance of data across multiple technologies, applications, and security domains, which are involved in their derivation.

We argue that provenance approaches developed in the context of databases and workflows, which are reviewed in Chapter 4, essentially deal with closed systems. By that, we mean that workflow or database management systems are in full control of the data they manage, and track their provenance within their own scope, but not beyond. A broader perspective is required by which elements of provenance information, captured by individual systems, can be brought together to describe the provenance of information flowing *across* systems. This is the specific purpose of the *Open Provenance Vision*, which we outline in this chapter, organised as follows.

First, we consider the architectural principles allowing a monolithic application to be made provenance-aware (Section 5.1), and then generalise them to multi-systems applications (Section 5.2). We then provide some background about the Provenance Challenge activity, a community project aiming at inter-operability of provenance technology (Section 5.3), which resulted in the Open Provenance Model described in Section 5.4. In Section 5.5, we then contrast open and closed world assumptions for provenance. We then review approaches that broaden their provenance tracking capabilities to other systems (Section 5.6).

5.1 Provenance-Aware Monolithic Application

The research community has now gained a fairly good understanding on how to make a single monolithic application *provenance-aware* [284], by this we mean an application that tracks the provenance of its data and allows for such provenance to be queried.

It is recognized by most communities (whether workflow, database, service oriented, or others) that extra information needs to be asserted and recorded as the application proceeds. In the case of databases, such information may be referred to as annotations [93] or simply provenance data [176]. In the case of workflow systems, it is also referred to as provenance information (e.g., Kepler [39, 41] and Taverna [419]); others refer to it as process documentation [201].

Without loss of generality, we will refer to the extra information to be captured as *process assertions*. *Process assertions are to electronic data what a record of ownership is to a work of art.* Provenance-aware applications create process assertions and store them in a *provenance store*, the role of which is to offer a long-term persistent, secure storage of process assertions (cf. Figure 5.1).

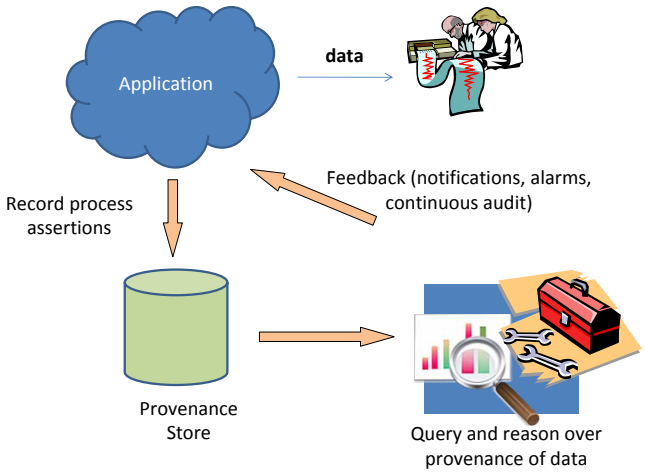


Figure 5.1: Provenance in a Single System

Once process assertions have been recorded, provenance can be retrieved and analysed by querying the provenance store. Analysis can vary from extractions identifying which source data were copied in a result (where provenance), to sophisticated rule-based checks to decide whether a process is compliant with a set of rules. Such checks can determine, for example, that source data are appropriately licensed, that computations are undertaken with the required precision, or that the process that was executed is following established practices. An exporting capability can be used to provide feedback to applications, by means of asynchronous notifications or alarms; continuous monitoring or audit functional-

ity can also be programmed.

Whilst Figure 5.1 depicts the provenance store as a separate entity, it may be integrated with application data, in separate tables in the same database (cf. PERM [176]). To date, there is no universal consensus on an internal format for process assertions, since it is often optimised for the technology used in the application. For instance, Woodruff and Stonebraker [406] propose to minimize the cost of storing process assertions by computing them lazily, as queries are issued to the provenance store, making use of an inverse function (for invertible functions).

5.2 Provenance Inter-Operability across Components

When data flows across multiple components, we could make each individual component provenance-aware by adopting the technique described in Section 5.1. However, there is a challenge to tracking provenance across multiple applications, since there is no common provenance model to describe the execution across multiple technologies, there is no agreed mechanism to connect the provenance of a received data item and the provenance of its matching sent data, and there is no query language and mechanism to operate over multiple provenance stores.

To address this challenge, the Open Provenance Vision is an approach that consists of controlled vocabulary, serialization formats and APIs (Application Programming Interfaces) that allow provenance from individual systems to be expressed, connected in a coherent fashion, and queried seamlessly. Specifically, the Open Provenance Vision promotes an inter-operability layer, based on the Open Provenance Model (OPM) [299], as illustrated in Figure 5.2.

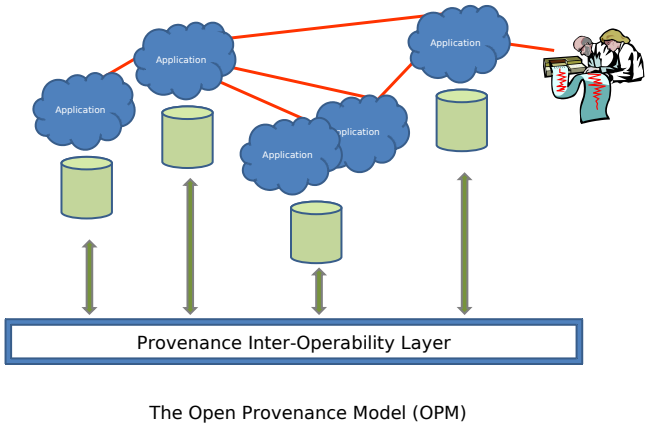


Figure 5.2: Provenance Across Systems

In Figure 5.2, we see a flow of information across multiple applications or systems. Each system is individually made provenance-aware, and making use of its own provenance store. To track the provenance of data produced by such an application, one would need to traverse the contents of all these provenance stores. Such a task is essentially impossible if all systems adopt their own provenance representation. Instead, an inter-operability layer is introduced, which allow the contents of individual stores to be exposed, and queries to be run uniformly across these stores.

OPM is a lingua franca for provenance systems, since it allows provenance to be represented in a technology agnostic manner and to be serialised in various formats such as RDF and XML. During the third Provenance Challenge (which we discuss in Section 5.3), it has been demonstrated that many systems can export provenance in the OPM representation, and that they can import it as well. A model such as OPM offers the ability to exchange provenance information in an inter-operable manner. From this model, we anticipate that query languages and APIs will be devised, and that query engines would be able to federate provenance information from multiple stores.

5.3 The Provenance Challenge Series

Over the years, a series of systems have been developed to track and exploit provenance in many different ways. Following a discussion session on standardisation at the *International Provenance and Annotation Workshop* (IPAW'06) [291, 34], a consensus began to emerge, whereby the provenance research community needed to understand better the capabilities of the different systems, the representations they used for provenance, their similarities, their differences, and the rationale that motivated their designs.

Hence, the first, second and third Provenance Challenges[1][2][3] were successively set up in order to provide a forum for the community to understand the capabilities of different provenance systems and the expressiveness of their provenance representations. In the first and second challenges, the participating teams ran an agreed Functional Magnetic Resonance Imaging workflow, exported provenance information, and implemented pre-identified "provenance queries" asking typical questions about past execution of the workflow [297]. Key themes related to the provenance challenge activity are summarised in the tage cloud of Figure 5.3.

As discussions indicated that there was substantial agreement on a core representation of provenance, the *Open Provenance Model* (OPM) [292] (subsequently revised by a broader committee [299]) was put forward as a data model by which systems can exchange provenance information. This model was the focus of a

[1] http://twiki.ipaw.info/bin/view/Challenge/FirstProvenanceChallenge
[2] http://twiki.ipaw.info/bin/view/Challenge/SecondProvenanceChallenge
[3] http://twiki.ipaw.info/bin/view/Challenge/ThirdProvenanceChallenge

Figure 5.3: Challenge Cluster Tag Cloud

Third Provenance Challenge, where its suitability was practically evaluated by using it as the agreed format for provenance information exchange.

From the outset, because precision matters when systems have to inter-operate, OPM was described in a technology-agnostic manner, both in natural language and using a formal notation. The key structure defined in the Open Provenance Model is an *OPM graph*, a directed acyclic graph aimed at representing data and control dependencies of past computations. The specification also outlined the kind of inferences that are permitted over such graphs.

5.4 The Open Provenance Model

The primary aim of OPM is to be able to represent how "things", whether digital data such as simulation results, physical objects such as cars, or immaterial entities such as decisions, came out to be in a given state, with a given set of characteristics, at a given moment. It is recognised that many of such "things" can be stateful: a car may be at various locations, it can contain different passengers, and it can have a tank full or empty; likewise, a file can contain different data at different moments of its existence. Hence, from the perspective of provenance, OPM introduces the concept of an *artifact* as an immutable piece of state; likewise, it introduces the concept of a *process* as actions resulting in new artifacts. The Open Provenance Model is a model of artifacts *in the past*, explaining how they *were* derived.

A process usually takes place in some context, which enables or facilitates its execution: examples of such contexts are varied and include a place where the process executes, an individual controlling the process, or an institution sponsoring the process. These entities are being referred to as *Agents*. They are a cause (like a catalyst) of a process taking place.

A provenance graph aims to capture the causal dependencies between the

abovementioned entities. Therefore, a provenance graph is defined as a directed graph, whose nodes are artifacts, processes and agents, and whose edges belong to one of following categories depicted in Figure 5.4. An edge represents a causal dependency, between its source, denoting the effect, and its destination, denoting the cause.

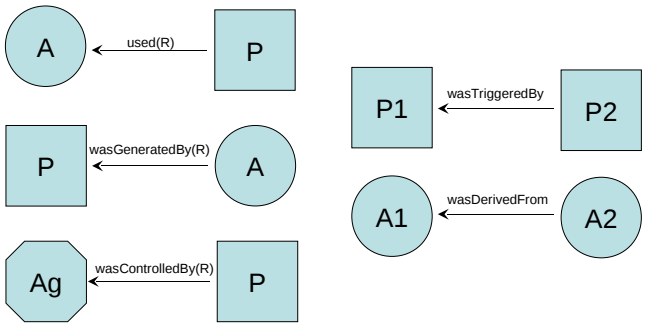


Figure 5.4: OPM Nodes and Edges

The first two edges express that a process *used* an artifact and that an artifact *was generated by* a process. Since a process may have used several artifacts, it is important to identify the *roles* under which these artifacts were used. (Roles are denoted by the letter R in Figure 5.4.) Likewise, a process may have generated many artifacts, and each would have a specific *role*. For instance, the division process uses two numbers, with roles dividend and divisor, and produces two numbers, with roles quotient and remainder.

A process is caused by an agent, essentially acting as a catalyst or controller: this causal dependency is expressed by the *was controlled by* edge. Given that a process may have been catalyzed by several agents, we also identify their roles as catalysts. We note that the dependency between an agent and a process represents a control relationship, and not a data derivation relationship. It is introduced in the model to more easily express how a user (or institution) controlled a process.

It is also recognized that we may not be aware of the process that generated some artifact A_2, but that artifact A_2 *was derived from* another artifact A_1. Likewise, we may not be aware of the exact artifact that a process P_2 used, but that there was some artifact generated by another process P_1. Process P_2 is then said to have been *triggered by* P_1. Edges *was derived from* and *was triggered by* are introduced, because they respectively allow dataflow or process oriented views of past executions to be adopted, according to the preference of system designers.

To illustrate the model, we revisit the application of Figure 3.1 and display in Figure 5.5 the provenance of the mashup, represented in OPM. (We note that

Figure 5.5 is a generalization of Figure 3.2, which only contained artifacts and *was derived from* edges.) At the bottom, we see the mashup artifact. Plain edges represent the *was derived from* relation, whereas dotted edges represent *used* and *was generated by* relations. Edges can be subtyped; the subtype is represented by a label alongside the arrow. In the middle of the graph, the User (represented by an Agent) selected tuple `tb2`, which was used to parametrise a query, that extracted copies of `p3` and `b2` to construct the mashup.

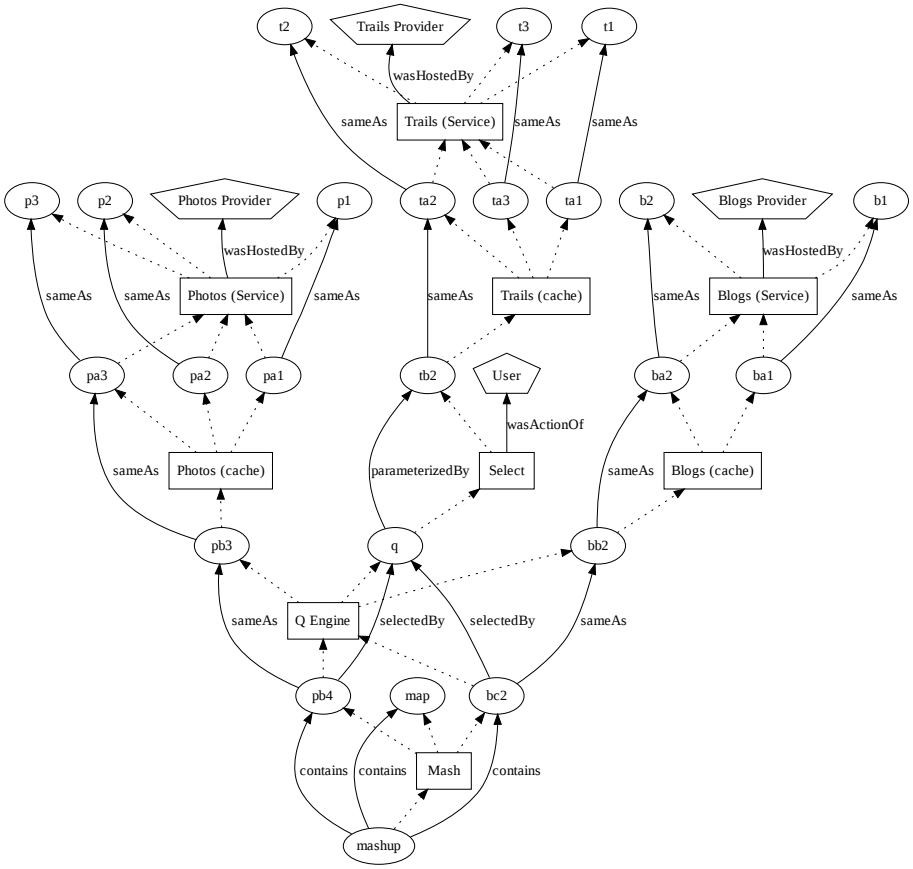
It is beyond the scope of this document to provide a complete description of the Open Provenance Model. Let us just mention a few salient features that characterize its expressiveness. OPM introduces the concept of account, which can be regarded as a graph colouring, identifying a graph subset containing a description of a past execution (by one or more witnesses). Multiple accounts can co-exist in a graph, and relationships such as overlaps, alternate or refinements are being defined. Within the scope of an account, the chain of *was derived from* edges is expected to be acyclic. Finally, OPM graphs can be enriched with optional time information, which is expected to be consistent with data derivation order.

5.5 Provenance in Open Systems

The Open Provenance Model was designed to represent the provenance of artifacts produced in open systems, by this we mean systems whose topology may not be known at design time, and whose components, location and identity may only be discovered at runtime. In such systems (even without assuming malicious behaviour), we should expect components to provide descriptions of execution that do not align exactly, since they are not omniscient and can only operate on the basis of their local observations and knowledge.

To allow for multiple descriptions of a same execution to co-exist in a same provenance graph, OPM's notion of an *account* can be seen as a colouring of graph that identifies a consistent subset of provenance information. Novel reasoning techniques are required to be able to reason with conflicting information provided by different accounts.

In an ideal world, provenance would be consistent and complete, providing the one and only one description of execution, for all activities and data, at all levels of granularity, involved in an execution. This is unfortunately not the case. Some components may not be provenance-aware, and may not be able to capture provenance. In that case, the components they interact with or wrappers may simply capture the interactions between them, without providing explanations of how data were internally derived inside the non-provenance component. Tan [380] categorized this kind of provenance as workflow and coarse-grained. We prefer to regard it as rather incomplete description in the continuum of possible provenance representations.



p1,p2,p3	original photo record
t1,t2,t3	original trail records
b1,b2	original blog records
pa$i$	copy of pi in cache
ta$i$	copy of ti in cache
ba$i$	copy of bi in cache
pb$i$	copy of pa$i$ retrieved by q
tb$i$	copy of ta$i$ parametrizing q
bb$i$	copy of ba$i$ retrieved by q
pb4	copy of pb3 returned by q
bc2	copy of bb2 returned by q

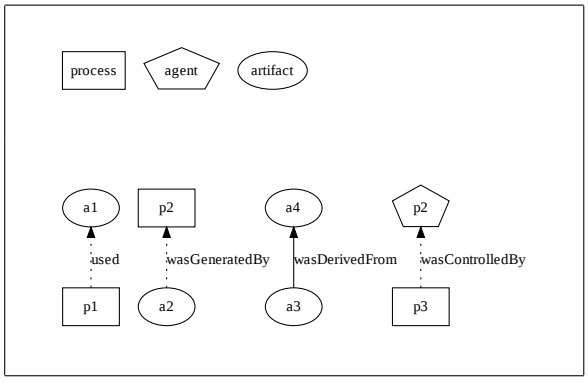


Figure 5.5: OPM Provenance of Mashup

We refer again to Re and Suciu's *approximate lineage* [332] as an alternative to complete provenance. Approximate lineage is *purposefully* designed to be incomplete, so as to compress provenance representation, while still representing the most influential facts of a derivation. In the Open Provenance Vision, however, a framework has been designed to capture as much process assertions as possible, but we acknowledge that, due to systems' imperfections, provenance may be *incomplete*. This view differs substantially from Gehani *et al.*'s [169], who assume complete provenance.

Caroll *et al.* [66] make an interesting observation on the open world assumptions of named graphs, which also apply to the Open Provenance Vision. In an open world, the provenance of an artifact is an open-ended collection of assertions about this artifact (similarly to RDF and OWL descriptions of a resource are considered to be open ended). Indeed, any component can always provide further information about a past execution, that is relevant to the provenance of an artifact. However, accounts can help structure such descriptions. Some approaches, such as PASOA [201, 204], allow a component to indicate that they have completed the set of assertions that have made about a given execution. Transposed to an OPM context, this would allow an account to be sealed, effectively, allowing a reasoner to infer that a faithful component has not observed some action (meaning either it did not occur, or that it could not be observed by the observer).

5.6 Broadening the Scope of Provenance beyond Closed Systems

Figure 5.2 illustrates that the Open Provenance Vision requires individual systems to collect their provenance, which can then be exported and integrated according to the OPM model. Chapter 4 has already discussed highlights of the literature regarding provenance in workflows and databases, which are typical technologies used in implementation of individual systems. In this section, we survey recent undertakings to make other technologies provenance-aware.

Bowers *et al.* [40] note that the scope of most scientific workflow systems (Triana, Taverna, Pegasus) is limited to single workflow runs. While VisTrails tracks workflow definitions, most implementations are largely ignorant of data management tasks carried out between workflow runs. By providing a structure to record project information and name collections, provenance can be traced across workflow executions. (Note that Kepler embeds provenance in the application data (as an output stream), and to some extent the proposed solution remedies some shortcomings of this approach).

In curated databases, the curators are an important "component" of the system since they make decisions to create, delete, annotate and edit database

records. Buneman *et al.* [51, 52] assume that changes to the database can be modeled using transactions comprising simple "copy-paste" updates. They derive a model of provenance that captures such update sequence, which allows them to reconstitute changes to the database. To achieve this vision, their implementation intercepts user's actions. Archer *et al.* [10] also tackle the curation problem but in the context of a "DataSpace". They introduce a history table that captures each user action with respect to a data relation. From it, they derive a notion of "provenance graph", which is a directed acyclic graph, where vertices correspond the current state of a data value of interest.

Given that many systems no longer have a native user interfaces, but instead offer a Web 2.0 rich interface accessible through the Web browser, Margo and Seltzer [264] investigate "browser provenance", in the context of which, they consider browser logs as process assertions, from which provenance can be derived.

Given that user interactions typically take place over their desktop (for users working at their workstation), it is also natural to consider provenance of information on the desktop. Shah [353] tackles this problem to provide a provenance-enabled search technique for files on the desktop. They use a binary rewriting technique to trace all file system and interprocess communication calls. They build a dependence tree based on the "kinship" relation, where a file is said to be an ancestor of another file, if the former may have played a role in the origin of latter. Ultimately, to be able to capture everything that occurs on the desktop, cooperation from the operating system would be required. In PASS [218], relevant kernel calls are trapped, and recorded by means of a logging interface, from which provenance logs are formed: at that level, the challenge is that the business logic cannot always be reconstituted; to address this problem, in addition to the kernel-level logging interface, explicit assertions can be made about application-level dependencies.

Capturing provenance on the desktop, in the browser, at the operating system level, or at the user interface inevitably brings ethics concerns: should we capture all user interactions with a computer, to the point that all information can be traceable? Privacy is an important concern, and any log must remain the property of the user. Therefore, the necessary security mechanisms need to be put in place to protect such information. Security techniques related to provenance will be discussed in Chapter 7.

5.7 Summary

The Open Provenance Vision is motivated by the architecture of today's information systems, which tend to be loosely-coupled dynamic assemblages of components, deployed in multiple security domains, and relying on multiple communication and execution technologies. The Open Provenance Model is the first attempt of a provenance representation, allowing cross-systems information flows

to be documented in a coherent manner.

Such a representation presents multiple research challenges, including those described in the previous chapters, such as design of query languages, optimised storage, account-aware inference techniques, and reasoning in the presence of partial information. A natural application of OPM is tracking the provenance of information over the Web, a topic, which we discuss in the following chapter.

Chapter 6

Provenance, the Web and the Semantic Web

The ultimate driver for the Open Provenance Vision described in Chapter 5 is the World Wide Web. The Web has become a global information space where the contents of databases are increasingly exposed directly. The Semantic Web facilitates the annotations of these data sets with RDF metadata, forming a global web of Linked Data [431]. Techniques such as mashups and RSS feeds integrate data from multiple sources, providing users with information customized for their needs. In this context, tracking provenance is perceived as a critical issue [212, 123] since it helps determine the quality of and trust one can put into data.

Issues in this area can be categorized in the following separate strands. *(i)* Given the importance of provenance, it is to be regarded as first-class data, itself to be exposed on the Web (Section 6.1). *(ii)* Semantic Web technologies are themselves being used, not only to represent provenance information, but also to query and reason over it (Section 6.2). *(iii)* Given the importance of metadata in the information discovery process, and the ease by which such metadata can be published on the Web, tracking the provenance of RDF-based information has also become a focus of investigation (Section 6.3). *(iv)* In the Semantic Web, not only can triples be asserted, but also they can be inferred. In such case, special techniques need to be devised to track their provenance (Section 6.4). Issues of data quality and trust, which are crucial over the Web, are investigated in Chapter 7.

6.1 Publishing Provenance on the Web

The principles of exposing information on the Web are now well understood [445], namely the use of Uniform Resource Identifiers (URIs) — a system for identifying resources globally — and protocols such as HTTP to access resources. Different

approaches for exposing provenance have been proposed in the literature, namely hypertext generation, RDF [446] views, Webdav [438], and REST [439], which we now discuss.

Two major constituents of provenance are identified by Zhao *et al.* [415]: annotations attached to objects (in a structured, semi-structured, or free text form), and derivations paths (from a workflow, query or program). In [415, 416], they describe a dynamically generated hypertext of provenance documents, data, services and workflows. Their aim is to support Hendler's vision of a Web of science [443]. This Web is created dynamically by means of ontology reasoning, annotation processing and link insertion.

SAM [306, 309], the scientific annotation Middleware (SAM) is a precursor system, pioneering the use of emerging Semantic Web technologies to separate the initial capture and storage of data and metadata from its subsequent presentation to others, hereby shifting the focus from up-front standardization to on-demand mapping of the data and metadata. In particular, SAM offers an electronic notebook capturing provenance of scientific experiment. By adopting the Webdav approach and URI identifiers, it allows navigation of provenance information. The pedigree browser allows for provenance browsing, and a portlet allows for graph visualization.

Hunter and Cheung [224, 91] presents Provenance Explorer, a system able to generate personalized views of the provenance relationships automatically using a combination of user input, semantic reasoning and access policies. Provenance information is extracted from a system that generates it (such as Kepler or Taverna), and inferences are made to build user views.

Moreau [447] identifies the key design rules for provenance architectures based on REST [439] guidelines. They include the identification of a provenance store by a URI, the ability to upload process assertions about execution to service by means of a POST request, the ability to identify artifacts by URIs, to query the provenance store by a single GET request, and the ability to retrieve a provenance query result by a single GET request. Retrieved provenance can be directly browsable (html format), in machine-processable formats such as XML or RDF, or ready for visualisation by the browser (JSON).

6.2 Semantic Web Techniques for Provenance

The use of Semantic Web technologies has been advocated to facilitate provenance acquisition, representation, and reasoning. On the one hand, RDF allows for resources to be referred to by URIs and its triple structure simplifies graph representation; the associated query language SPARQL [450] easily expresses their querying. Finally, OWL can be used for ontological definitions and reasoning. The tag cloud of Figure 6.1, produced from papers of the bibliography with a focus on Semantic Web techniques, identifies key issues in this area.

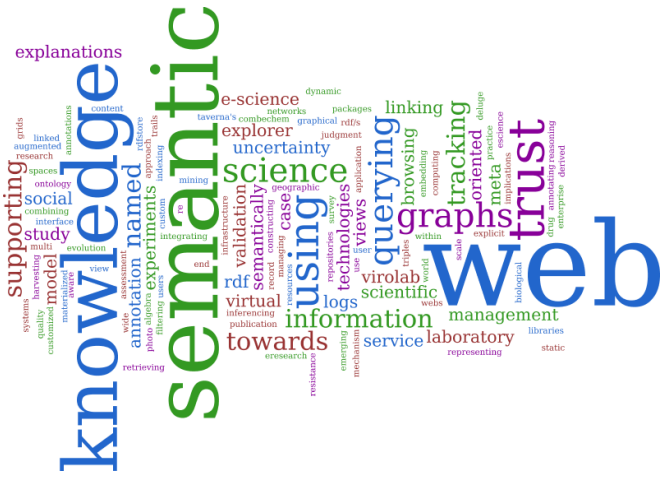


Figure 6.1: Semantic Web Tag Cloud

Zhao *et al.* [416] view provenance information from four different levels: organisational (incorporates who ran the workflow), process (a kind of event log, capturing inputs, outputs), data (captures data derivation) and knowledge (annotation about all the previous). Zhao *et al.* [419, 421] advocate the use of RDF [446] to represent provenance information, LSIDs [1] to identify resources, and ontologies to deliver a common semantic view of such data. Such a Web of data can be visualised by a Semantic Web visualization tool, or can simply be navigated with a browser.

Chen *et al.* [81, 79, 78] use the term "augmented provenance" to denote the provenance of a piece of data and related semantic metadata about the process that led to such a piece of data. They explain how such semantic metadata can be captured from the workflow construction environment and the workflow enactment engine.

Sahoo *et al.* [343, 342] use the term "semantic provenance" to denote provenance in which domain knowledge and ontological underpinning have been incorporated. Like Zhao *et al.* [415], they advocate a semantic service, which incorporates domain specific knowledge into representations. It is an approach that is adopted by several systems, by means of specialising their data dependenci for specific application domains.

Myers *et al.* [308] observe that the disconnect between processes and data, where scientists have to manually operate heterogenerous tools with little integration, preventing experiments to be reproduced easily, and the loss of the collaborative contexts (notes, discussions, emails) are such that by the time results are published, most traces of the original process and data are inaccessible to the reader. To address this concern, they advocate the use of a semantic content

[1] lsids.sourceforce.net

management system, of which Tupelo is a core constituent. Tupelo [308, 397] is a middleware that provides a Web access protocol and Java API that interface with an RDF mapping of the Open Provenance Model. For this system, Wang *et al.* [397] propose a specialisation of OPM for GIS applications.

Golbeck *et al.* [182] demonstrate the flexibility of Semantic Web technologies to implement the Provenance Challenge. A provenance ontology was specified in OWL, and the description of the execution expressed accordingly. SPARQL was then used to implement the challenge queries. While the approach was demonstrated to work with another system, it makes some strong assumptions about application data, which typically remain in the control of the application.

Miles *et al.* [286, 404] use the OWL reasoning capabilities to determine the semantic validity of past experiments. Subsumption is used to determine that the actual inputs and outputs of past experiments have the expected types; actual performed operations are checked to be conformant to a plan; legal constraints associated with input data sets are checked to verify patentability of results.

Halaschek-Wiener *et al.* [209] present a Semantic Web portal to annotate digital images and track their provenance, in the form of annotations such as submitter name and email. Provenance is browsable and actively used to enrich the user's browsing experience.

In the context of a Chemistry lab, Frey *et al.* [162] envisage an RDF-based semantically-described world, in which a policy of "annotation at source" is enforced to track all information, including provenance, of manipulated digital and physical artifacts. Whenever data is processed, it is annotated with a description of the processing, effectively making its provenance explicit.

In the context of a virology application, Balis *et al.* [13, 12, 14, 11] propose a Query Translation Tool (QUaTRo) which allows users to construct queries to underlying provenance and data repositories with wizards and using technology-independent concepts, expressed in the terms of the domain familiar to end users.

6.3 Provenance for RDF

Whilst many authors advocate the use of Semantic Web technologies to represent and query provenance, Carroll *et al.* [66] take the opposite view, and identify the problem of provenance of triples (and other issues such as versioning and signature) in RDF. They propose named graphs as an entity denoting a collection of triples, which can be annotated with relevant provenance information. The RDF triple is the atomic assertion permitted on the Semantic Web; attaching authorship and origin to such assertions is inline with the PASOA approach [195, 377] which considers collections of assertions, cryptographically signed, as the foundation of provenance information. The named graph proposal follows a series of approaches to address the problem of provenance and signature of RDF triples [333, 339, 234], where triples are extended with some construct allowing

provenance to be expressed. None of these approaches however specifies how provenance itself should be represented, but they simply offer a placeholder for its representation.

Pediaditis *et al.* [319] argue that named graphs alone cannot capture provenance information in the presence of rdfs reasoning and updates. Given two triples belonging to separate named graphs, which graph does any tuple inferred from these belong to? In other words, shared origin cannot be captured by named graphs alone. To remedy this problem, they propose a new construct, a graph-set, which allows them to capture and query provenance information adequately. This construct bears a strong similarity with the notion of account in OPM [299] since OPM edges may be asserted to belong to multiple accounts, meaning they belong to different descriptions (potentially from different observers).

Watkins *et al.* [399] explain how named graphs allow them to define a partitioning of their graphs, which then can be signed, effectively creating a Warrant Graph [66], and track the provenance of documents in a software version control repository.

Zhao *et al.* [420] envision subject-specific data webs that integrate multiple sources of scientific data (published in separate databases) in a seamlessly integrated view across the web. To allow scientists to maintain their trust across this web of independently evolving databases, provenance metadata is produced. It relies on RDF named graphs, over which they provide evidence for links and traces of how links are updated and maintained.

Gibson *et al.* [172] propose an approach that leverages named graphs and extensions to the SPARQL query language to create and manage views as a server-side function, effectively customising the presentation of provenance to users. With their approach, multiple operations can be aggregated in a single operation, hereby hiding details that are not considered important to the user. The notion of account and associated refinement relation in OPM are a mechanism that offers a similar form of abstraction.

Futrelle [163] also proposes that attribution and timing information for each triple be represented using Dublin Core creator and date properties, using an actor URI for the value of the creator element and an ISO 8601 timestamp for the value of the date element. Ding *et al* [123] introduce the notion of RDF molecule, an RDF graph partition, which offers a level of granularity between graph and triple. The use of molecules is demonstrated to track provenance of a graph.

Hartig [212] proposes a specialisation of OPM, referred to as *provenance vocabulary*, to describe the provenance of Linked Data over the Web. His model accounts for the creation and access of RDF data.

Dividino *et al.* [125, 126] focus on the problem of querying data and at the same time querying associated meta-knowledge such as provenance, authorship, recency or certainty of data. Their approach consists of meta-knowledge in RDF, and specifically represents provenance using an RDF serialisation of HOW-

provenance [188]. Their proposed query language is an extension of SPARQL that allows meta-knowledge queries also to be expressed.

6.4 Knowledge and Web Provenance

McGuinness and Pinheiro da Silva introduce Inference Web [270, 271, 113] an extension of the Semantic Web which aims to enable applications to generate portable explanations for any of their answers. A key component is PML (Proof Markup Language) which includes support for Knowledge Provenance and Reasoning information; PML includes metadata such as authorship and authoritativeness of a source, but also reasoner assumptions (e.g. closed vs open world, naming assumption) and a detailed trace of inference rules applied (with variable bindings). Relationships capture notions of Consequent and Antecedents to a proof step, the succession of which consists of a proof. Human-readable explanations are derived from the proof markup language, and browsable representations can also be exposed to the user. PML is shown to be convertible to OPM representations in the third Provenance Challenge[2].

Rio *et al.* [337, 336] describe how the Inference Web's knowledge provenance can be used to semantically annotate maps and how this semantic information can help scientists understand and evaluate map products.

Fox and Huang [152] also adopt the term Knowledge Provenance (KP), to address the problem of how to determine the validity and origin of Web information by means of modelling and maintaining information sources, information dependencies, and trust structures. They argue that given that the Web will always be a morass of uncertain and incomplete information, it is possible to annotate Web content to create islands of certainty. Knowledge Provenance consists four levels of provenance that range from strong provenance (corresponding to high certainty) to weak provenance (corresponding to high uncertainty). Static KP (level 1) focuses on provenance of static and certain information. Dynamic KP (level 2) considers how validity of information may change over time. Uncertainty-oriented KP (level 3) considers uncertain truth value and uncertain trust relationships. Judgement-based KP (level 4) focus on social processes to support KP.

Gomez-Perez and Corcho [184] investigate the use of Problem-Solving Methods as a mechanism to extract higher-level knowledge-oriented provenance from existing provenance traces. Such higher-level provenance is intended to be easier to understand by users, and allows them to better grasp vast amount of provenance information.

[2] http://twiki.ipaw.info/bin/view/Challenge/TetherlessPC3

6.5 Summary

As information dynamically flows across the Web, users need to have reliable means to obtain its provenance to decide whether they can trust information they access. Furthermore, the emerging sets of Linked Data, a pragmatic route to a semantic Web, form a network of pointers that allows automated navigation to data that is relevant to the user. In this context, reasoners need explicit representations of provenance in order to make trust judgements about the information they use.

This chapter has shown how useful Web and Semantic Web technologies can be exploited to represent, make accessible, query and reason over provenance information. Semantic Web technologies themselves are also susceptible to a provenance problem. Techniques are emerging to express both the relevant authorship of atomic assertions on the Semantic Web, and detailed reasoning process when these items of knowledge are being inferred by reasoners. This latter work, which has mostly been developed independently of provenance research in the workflow community, has recently been demonstrated to be compatible with it, in the third Provenance Challenge. All these elements are indicative of a convergence towards the Open Provenance Vision for the Web.

Chapter 7

Accountability

Complex organisations and systems are typically formed by assembling multiple autonomous entities that agree to cooperate in order to achieve overarching objectives. Such assembly of autonomous entities is often regulated by constraints in the form of norms, contracts or policies, specifying the responsibilities of entities, their obligations, permissions, and penalties incurred when failing to deliver.

This approach to organising complex systems existed well before the prevailing use of the Web, yet the pervasive use of the Web offers new opportunities for creating such complex organisations, quickly, dynamically, and for negotiating the rules governing them on the fly. For example, virtual organisations [441], in which autonomous agents collaborate to deliver composite services, provide a means of exploiting such possibilities.

However, this presents end-users with challenges, since they are now confronted with a very dynamic and fluid environment, where it is difficult to understand which entity is responsible and accountable for which action. Here, the term "end-user" must be understood in its broadest sense: end-users may be organisation customers, the organisation's participants, or even regulatory authorities.

Given that such applications are formed by assembling components dynamically, static methods that analyse their source code to infer their properties are not suitable. Even in systems that seemed to have substantial design and analysis before deployment, public cases of breach typically require an investigation, *after the fact*, to understand the origin of the problem[1]. This strongly suggests the need for approaches where systems faithfully document their execution, for potential future investigation.

Weitzner *et al.* [402] argue that, for information, "accountability must become a primary means through which society addresses appropriate use". For them, "information accountability means the use of information should be transparent so it is possible to determine whether a particular use is appropriate under a

[1] http://news.bbc.co.uk/1/hi/uk/7103911.stm

given set of rules, and that the system enables individuals and institutions to be held accountable for misuse". Dynamically assembled systems need to be made *accountable* for users to gain confidence in them, i.e., their past executions must be auditable so that compliance with, or violations of, policies can be asserted and explained.

Weitzner *et al.* [402] note the similarity between accountability and provenance in scientific experiments. We see provenance as a key enabler for accountable systems since it consists of an explicit representation of past processes, which allows us to trace the origin of data, actions and decisions (whether automated or human-driven). It therefore provides the necessary logs to reason about compliance or violations. In this chapter, we review *the use of provenance to make systems accountable, or to derive trust in results and systems*. Given the critical importance of provenance in such contexts, the audit results or derived trust will be reliable only if provenance has been faithfully stored and has not been tampered with. We therefore begin with a review of security in the context of provenance (Section 7.1). We discuss novel approaches to check compliance, and make systems accountable, using provenance (Section 7.2). We then survey various approaches that have emerged to determine quality and trust of data, based on provenance (Section 7.3).

7.1 Provenance and Security

Braun *et al.* [45] make the case that securing access to provenance is different to "traditional data". First, given that provenance contains relationship between entities (such as artifacts, processes and agents), each relationship reveals information about the parties in the relationship, and therefore needs careful handling. Second, a data product and its provenance may have different sensitivity. In an employee's performance review, the data product is the review itself, which is available to the employee; its provenance encompasses the authors of the review, who have to remain anonymous. Hence, the employee can see the data but not its provenance. Symmetrically, a University applicant typically provides the names of the referees (and sometimes the actual reference in a sealed enveloped); the reference is to remain invisible to the applicant, while its provenance is known to the applicant.

Figure 7.1 summarises the key themes pertaining to security and provenance, they relate to access control, provenance integrity, non-repudiation of provenance, and sensitivity of provenance information. We discuss these themes in the following sections.

Figure 7.1: Security Tag Cloud

7.1.1 Access Control

In many systems or approaches, it is often considered that all users can see all provenance information. Yet, as argued by Braun *et al.* [45], this is not realistic, and hence access control needs to be introduced to repositories of provenance data [31].

Following the spirit of Braun's distinction of data access and provenance access, Chong [94] revisits Cheney's traces in the context of secure provenance. He considers two kinds of policies for application data (confidential or public) and two kinds of policies for the associated provenance (confidential or public). A location is said to have confidential provenance if it should not be known whether this location affects the result of the computation. He proposes extensional security requirements that restrict what information is revealed by the provenance trace, and that ensure the program result respects the provenance security policies.

Rosenthal *et al.* [338] make the case that role-based access control is not easily extended to support the security requirements of multiple stakeholders related to a given provenance trace. They propose to structure distinct concerns in a modular fashion to facilitate maintenance: namely, security, legally mandated privacy, and organizationally mandated privacy. They annotate OPM entities with access control attributes. They promote the use of ABAC (attribute-based access control) over RBAC (role-based access control), as the latter suffers from scalability problems, when policy becomes finer-grained and more attributes are involved, and roles have to be created for each combination of attribute, making the management of user to role mapping challenging.

Syalim [374] considers different levels of granularity of access control over provenance information. Policies can be associated with artifacts, processes, or paths of dependencies. Access rights are granted to groups of users.

Chebotko *et al.* [76] extends user views [98, 30] with security considerations: specifically, the notion of security view consists of a portion of a provenance graph for a given user, according to the access control policy prescribed by the user's role. Access control policies are specified by workflow designers (in terms of the workflow building blocks (task, port, data channel), and inherited by the derived provenance produced at execution time.

Nagappan *et al.* [310] investigate the problem of confidential provenance in the context of Kepler; they introduce the notion of query sharing, by which users can explicitly share queries over provenance with their collaborators.

7.1.2 Provenance Integrity

Provenance vouches for the origin and authenticity of the data it relates to. For such a guarantee to hold, provenance itself must be preserved in its original form without any falsification or tampering. Integrity, in Information Technology terms, means that data remains unchanged while stored or transmitted.

Hasan *et al.* [215] are concerned with undetected rewrites of history, which occur when malicious entities forge provenance chains, in order to fake the authenticity of a document or data set. They consider provenance of a document as a linear chain of the principals performing actions on that document. They target chain forgeries that maliciously add new chain entries and make after-the-fact modifications, and offer the following integrity assurances:

- An adversary acting alone cannot selectively remove other principals' entries from, or add entries in the chain, without being detected by an auditor.
- Two colluding adversaries cannot add entries of other non-colluding users "between" them in the chain without being detected by the next audit.
- Once the chain contains subsequent entries by non-malicious parties, two colluding adversaries cannot selectively remove entries associated with other non-colluding users between them in the chain, without being detected by the next audit.

Their solution consists of propagating cryptographic checksums along the chain, allowing entries to be sequentially validated.

Factor *et al.* [135] consider the problem of long-term archiving of data, and note that in most cases, digital objects cannot be preserved without any change in the bit stream, and that we have to modify the original object to have the ability to reproduce it in the future. This leads to a paradox since preservation entails change, while authenticity needs fixity. To address this concern, they rely on provenance to track changes that occur to data during the preservation activities, and they preserve provenance alongside data.

Gadelha [166] considers some security properties of provenance — integrity, confidentiality and availability — in the context of intellectual property conflict claim resolution and of the reliable chronological reconstitution of scientific experiments. To address these, they secure data authorship and temporal information of provenance records. To this end, they use the time-stamping protocol in addition to cryptographic signatures. So, given a raw provenance record, they digitally sign the record, compute a hash value, which they send to the time-stamping server, which signs the hash and the current date.

Gehani and Lindqvist's aim [168] is to reliably determine the lineage of a piece of data. Their model of provenance is one in which "lineage metadata" is communicated along application data; their concern is lineage space requirements increase, as data get processed. Their solution is not to relay provenance but to leave lineage details at the nodes where the operations occur and forward cryptographic commitment to prevent repudiation. Their solution [169, 170] provides operation chain non-repudiation since the signed lineage of each input is extracted and added to the metadata of each output, itself, being hashed and signed.

Zhang *et al.* [412] define the pedigree forgery attack, as the situation where an attacker presents to the data recipient a predigree and data product, such that the pedigree does not accurately describe the data product's authorship. To prevent an attack, they introduce a cryptographic proof in the audit records so that to the data recipient can check that the pedigree associated with a data product is correct. By a cryptographic protocol, they ensure that the output of a workflow step matches the inputs of its successor. Their approach is not as strong as chains discussed by Hasan *et al.* [215] since only input/output matches are protected and not the overall chain.

7.1.3 Liability and Accountability for Provenance

An important consideration in any provenance system is the accuracy or objectivity of the assertions recorded [377]. Most systems capture statements about some aspect of a process by some of its components. From a more abstract viewpoint, such statements are however only a subjective view of that aspect by a component. It can be difficult sometimes, if not impossible, to determine how closely this view tallies with actual reality. Therefore, it becomes paramount to forge a clear link between a component and an assertion that it is responsible for. Such a link, which can be provided through digital signatures [377, 195], ensures that responsibility and corresponding liability is attributable to the correct component.

Cryptographic signatures let one determine the source of a metadata assertion [257] (or, more precisely and more generally) the identity of the system, person or organization that stands behind the assertion, and to establish a level of trust in this identity. One can have near-absolute confidence that the source

possessed the requisite public/private key pair (assuming that the private key has not been compromised, and the key pair has not been revoked); the level of trust is in the binding of identity to possession of the key pair.

Cryptographic signatures, as we have seen in this section, address several properties of provenance security: non-repudiation and liability for the actions performed by components, authentication of the individual assertions, and finally integrity and unforgeability of the assertions.

7.1.4 Sensitivity of Provenance Information

In a basic example, the assertions pertaining to a message exchange between two components would simply contain the contents of that message verbatim. Depending on application domain requirements, however, parts of the message may need to be obscured or transformed in some way when they appear in a provenance record. A good example of this is found in the electronic health care records domain [230], where privacy requirements mandate that patient identity on health care records be anonymized if the information on the record is being utilized for non-diagnostic reasons (for example, to answer statistical questions about medical processes).

If provenance is utilised in such a context, then certain data items (such as patient identifiers) that are transmitted in clear text in the original message exchange between actors must be obfuscated in some manner when stored as part of any provenance record. An approach, referred to as documentation style [195] has been proposed to address this problem: it applies a transformation to application messages before storing them in a provenance store.

7.2 Accountability

Figure 7.2 displays the tag cloud that summarises the key issues pertaining to accountability and provenance. In this context, a system is accountable if it can provide explanations for its actions, if its past actions are accountable, and if it can be demonstrated that its processes and decisions are compatible with rules, policies, or broadly regulations. The tenet of accountability is to keep a detailed record of past activities explaining how every data item is derived, and what triggered every action: in other words, with explicit representation of provenance, one can make systems accountable: provenance provides the necessary *evidence* which makes systems *transparent* and allows an auditor to determine whether *policies* are satisfied.

Groth *et al.* [202] derive, from philosophy and history, several principles that documentation of the past, in the form of process assertions, should implement, so that it can be seen as a proper evidence of past execution. *(i)* Assertions must be based on *observations* made by software components, according to the

Figure 7.2: Accountability Tag Cloud

principle of data locality; *(ii)* Each assertion making up documentation of the past for a computer system must be *attributable* to a particular software component. Of course, in some cases, it is reasonable to expect a component may make inferences about the world, and events that are not directly observable, but such inferences have to be marked as such, so that an auditor can distinguish observation from inferences (or guesses).

Aldeco and Moreau [2] propose a provenance-based architecture for an accountable system, and apply it to the data protection act, the UK implementation of the EU directive on private data. The architecture identifies multiple roles, such as data subject (the owner of private data), data controller and processor (those that manage and process private data), and the auditor, who aims to determine compliance of the system. Furthermore, the architecture relies on a provenance store acting as a trusted, secure and persistent repository of evidence, which the auditor can trawl to verify compliance.

Curbera *et al.* [108] distinguish business provenance from business activity monitoring (BAM). The latter is mostly focused on real-time access to business performance indicators, including interactive and real-time dashboards and proactive alert generation, whereas the former adds a historical perspective to BAM that enables root cause analysis and process discovery. With a focus on business compliance, they consider a typical business workflows, which differs from scientific workflows, since documents are exchanged by emails, published on and downloaded from the Web, and submitted to specific applications. Their model allows for data derivations to be inferred, by matching identities of business artifacts stored in a central provenance store.

Miles *et al.* [281] consider the problem of contract violation, i.e., where the responsibilities of the different parties and expectations are not met. They rec-

ognize that long-term business relationships require flexibility in the face of *mitigating circumstances*, i.e. assumptions that do not conform to those of the contract. Their purpose is to design a system that can handle such mitigating circumstances automatically. To do so, they require a reliable documentation of what has occurred and how it caused a violation. Determining the cause of the violation is performed by executing provenance queries; after checking whether mitigating circumstances prevails, they apply the appropriate policy to handle the violation.

Likewise, Vazquez-Salceda and Avarez-Napagao [391] consider the problem of runtime governance of service-oriented architectures. Their approach relies on detecting violation states that agents may enter into, and the definition of the sanctions that are related to the violations. An enforcement component relies on a rule engine to reason about the evidence stored in a provenance store to take decisions and plan actions whenever violations are observed.

Philip *et al.* [320] argue that provenance could be used in e-Social science, to track evidence-conclusion chains, so that decisions can be justified in terms of evidence and applied reasoning. They discuss philosophical, ethical and legal issues that such an approach would raise. Chorley *et al.* [95] discuss a representation of provenance and its application to evidence-based policy assessment.

In the context of copyright management, Ockerbloom [312] notes that to reliably determine the rights to a work, one may have to understand and record the provenance of a work, the provenance of its rights, and the provenance of the information used in rights determination. The factual assertion chains can be complex in their structure, and involve varying degrees of uncertainty. Evaluating the reliability of such assertion chains and reasoning with incomplete provenance are important issues to be considered by the provenance community.

7.3 Data Quality and Trust

As users delegate important tasks to systems and endow them with private data, it is crucial that they can put their trust in such systems. Accountability as defined previously is a way by which trust can be built, since action transparency and audit help users gain trust in systems. However, users may not always want (or have the resources) to audit systems; instead, they would like to be given a measure of trust, which they can rely upon to decide whether to use a system or not. The topic of trust has been extensively reviewed [181, 442, 451, 453]. Trust is usually based on an agent's own experience with respect to past interactions with other agents, whereas reputation draws upon information gathered from third-parties. In this section, we review work that derives a notion of trust in data from the provenance of data. We note that in some context, quality of data can be similarly derived from its provenance. Given a method to compute trust in data, it may then become possible to derive trust into systems by "aggregating"

trust about all the data they produce.

Golbeck reviews trust issues on the World Wide Web [181] and identifies provenance as a key element necessary to derive trust. Golbeck *et al.* [180] propose a trust inference algorithm that operates over Semantic Web data. It is applied to the Friend of a Friend ontology, used in social networking [183]. Likewise, Harth *et al.* [211] also argue for the social provenance of data, identifying the people or groups of people who originated data.

One important issue in determining data integrity is the trusworthiness of source providers and intermediate agents [114]. Dai *et al.* [114] propose some recursive functions that computes trust scores for data, depending on the trust of the information used to generate it and the trustworthiness of parties that handle it. They rely on a very simple generation path with a linear topology. Whilst this seems to be a step in the right direction, this work suffers from some limitation, such as the nature of the computation being performed; more complex topologies, and potentially multiple accounts of execution should be considered to make a system usable in practice.

Rajbhandari *et al.* [323, 324, 328] use provenance information to evaluate whether an abstract workflow description has been adhered to, and to enable a user executing a workflow-based application to establish trust in the outcome of a physical workflow. Their notion of trust is compositionally derived from the trust in processes (i.e. abstract workflow), the trust in services (i.e. physical workflow) and trust in source and intermediate data. A decision tree is used to decide whether a final data product can be trusted or not. Their original model offers a binary outcome (trusted or not) and was then extended to allow for a range of trust values [325].

Prat and Madnick [321] argue that believability is an essential characteristic of data quality and provide a precise approach to compute its measure using provenance. The measure is structured in terms of several building blocks: metrics for assessing the believability of data sources, metrics for assessing the believability of data resulting from process execution, and global assessment of believability.

Hartig and Zhao [213] define a notion of quality derived from the first author's Web provenance vocabulary [212]. They apply their approach to timeliness of data.

When a contract has to be drawn with a new business partner, there may be very little prior interaction with the partner (and possibly no reputation) to derive a notion of trust. Hence, one needs the means to decide whether to trust the contract itself. Groth *et al.* [199] propose an approach to measuring the success of prior contract executions, and a notion of contract similarity, which they use to determine the trustworthiness of contract proposals.

7.4 Summary

Information transparency is a desirable design principle for next generation Web-based information systems. By giving users the means to understand how information is produced and used, systems can be made accountable for their actions and the information they offer. Provenance is therefore a crucial technology that can provide such a transparency. It will be effective in achieving such a goal only if it cannot be forged and its integrity can be ascertained. Cryptographic techniques effectively applied to provenance can ensure such strong properties.

We have reviewed several approaches that adopt provenance as the foundational layer of accountable systems, which allow their actions and information flows to be audited, and their compliance or violation to rules and policies to be determined. Such strong capabilities — namely, information transparency, auditing capabilities, and compliance detection — provide users with the means to decide whether they can trust systems and information.

In practice, such a notion of trust needs to be derived for users, since they do not have the skills, time and will to audit systems and review their information flows. Several approaches have been developed to infer measures of trust, quality or believability from provenance information.

Chapter 8

Conclusion

In this article, we have postulated that "society can and should reliably track and exploit the provenance of information on the Web". By means of a journey through the provenance literature, we have developed arguments to establish this thesis. We summarise them as follows.

1. By means of blogs, social networking, news feeds, instant messaging, and collaborative tools, the Web has become a global and universal communication medium, exploited by businesses, governments and individuals. Concerns of privacy and security are being inevitably raised, as our lives become more dependent on the Web. To address these, information and systems available over the Web can be made accountable by introducing information transparency, auditing capabilities, and compliance checking tools; in turn, accountability will allow trust networks to be developed. In Chapter 7, we have argued that provenance provides solid foundations for building accountable systems, and with the appropriate security techniques to ensure its security, provenance can help establish the quality of data.

2. However, provenance is non-existent in today's Web applications, but is emerging in niche areas. In Chapter 6, we have reviewed state-of-the-art efforts to make provenance accessible on the Web, and integrate with Semantic Web technologies underpinning the Linked Data effort.

3. Given that information flows across multiple services over the Web, being transformed, filtered, processed and repackaged in many different ways, a representation of provenance has to be assembled by bringing evidence of local transformations and derivations into a coherent whole. This is the purpose of the Open Provenance Vision, and its community-driven Open Provenance Model, which we have discussed in Chapter 5. Integrating the Open Provenance Vision with the Web architecture is therefore a critical step in ensuring that provenance of Web data can be tracked and used.

4. For information provenance to be traceable over the Web, each information system or service involved in a global information flow has to track provenance in its local activities. Given that the majority of research has been conducted by the workflow and database communities, Chapter 4 has surveyed the key issues these communities have investigated. A point worth noting is that attention has recently been given to systems that involve humans who affect decisions and information flows, and therefore are entities belonging to their history.

5. As research is being undertaken by different communities, multiple notions of provenance have emerged (even within a single community). In Chapter 3, we have introduced a general definition of provenance, which was shown to be compatible with the prevailing definitions, and could suitably serve as a definition for the origin of information over the Web.

6. Using the Citebase tool, we have analysed an extensive bibliographical database on provenance, and identified research fronts, in Chapter 2. These research fronts, which take the form of co-citation clusters, concern topics that are aligned with the argumentation developed in this article: by means of research in databases and workflows, in the Open Provenance Model and the Provenance Challenge activity, in Semantic Web, and in security, the research community has already begun laying out the foundations for provenance on the Web.

8.1 The Benefits of Provenance on the Web

Having built the case for the thesis "society can and should reliably track and exploit the provenance of information on the Web", we now revisit some of the drivers for provenance discussed in Section 1.1, and recast them in a Web context.

After discovering a workflow close to their needs, a scientist iteratively adapts and modifies it for their experiment, which relies on several data sets, imported from highly curated databases, and other online data sets available from their research colleagues. When satisfied with the latest run of the workflow, the scientist makes all the data available over the Web, with their provenance, which has been captured by the workflow engine and all invoked libraries and services. The edited workflow is also published on the Web, with it provenance, including credits that can be automatically generated from its provenance. The scientist writes an article about their experiment, which includes a series of plots (not dissimilar to the one of Figure 2.1).

A journal publishes the article online, including its full provenance, with reference the data sets used in the experiment and the workflow; the plots contained in the article have also their provenance on the Web, and it is possible to verify how they were produced, with which data sets. Since the provenance of any data item

published over the Web is now accessible, a series of services have been deployed for the scientific community. The service `science-replay.org` is capable of replaying executions to verify results, or replay the workflows against new data sets. The service `science-quality.org` analyses the actual flow of execution that was involved in the generation of data sets; from this, it infers error propagation and a measure of the quality of data. The service `license-comply.org` identifies all initial data sets, workflows, services and libraries invoked in an experiment, and checks their licensing conditions are respectively satisfied. Overall, by making provenance of all data available over the Web, the scientific process is being made more transparent, scientific results can be better verified, and reused; provenance is therefore a key enabler of Hendler's vision of a Web of science [443].

As we have indicated, provenance is useful well beyond the world of science. Consider a shopper who finds some nice vine tomatoes in a supermarket, with a label indicating they were produced at a local farm. Using their mobile phone, they access the online information about that product, including its provenance. The supermarket, and its supply chain, including the transport company and the farm, believe that open-ness gives them a competitive edge, and therefore, expose all provenance information on the Web. Independent online services can exploit this information: `www.yourcarbonfootprint.com` computes the carbon footprint of these tomatoes and its actual food miles; `e-organic.com` is a service that provides an independent measure of the organic nature of products whose provenance is available on line. (This measure itself is available online, and its provenance can be obtained and audited.) With explicit provenance, shoppers could be offered a wealth of services that help them assert the quality of the product they buy (online or offline). Of course, realistically, we cannot expect shoppers to undertake such extensive analysis themselves, for every item they buy, especially if low price. Instead, they may rely on online shopping assistants making recommendations for them, according to their preference.

8.2 Future Research

This survey has identified key building blocks that would be necessary to accomplish our vision of tracking and exploiting provenance of information on the Web. However this vision is by no means implemented, and we are confronted to multiple conceptual and practical research challenges:

- Foundations: how can semantic frameworks and associated definitions be extended to a global information space such as the Web.
- Representation and Architecture: the Web is about standards, and these need to be agreed to represent, record and query provenance.
- Systems: systems need to be built to cope with the scale of information involved in this endeavour.

- Humans in the loop: non-intrusive methods need to be developed to better capture user's actions and reasons for performing actions, and visualisations techniques for provenance need to be devised for users to easily understand and navigate such information.
- Reasoning: novel techniques need to be devised to deal with provenance that is incomplete, conflicting, or not authoritative.
- Accountability: new services need to be conceived that can exploit provenance to offer auditing capabilities, compliance checks, and ultimately help users decide whether they can trust information over the Web.

Importantly, there is also a human challenge to achieving such a technical vision. While it is highly desirable to understand the origin of decisions and Web information, privacy is becoming an issue if all our actions on a computer are monitored and archived. The Web is recent and is being used as a quick dissemination tool. However, to make the Web provenance-aware, mentalities have to change: it is no longer sufficient to publish data, but associated provenance must also be made available. While tools may assist in this task, this inevitably increases the human effort involved. Hence, the cost-benefit of publishing provenance needs to be analysed. We believe however that when there is a strong requirement for accountable information, the benefits of provenance largely outweigh the cost of tracking and maintaining it.

Acknowledgements

Thank you to Chaomei Chen for his help with CiteSpace, to Danius Michaelides for his help with scripts for processing bibliographical data, and Ewa Deelman, Paul Groth and Simon Miles for providing feedback on this survey.

Provenance Bibliography

[1] Parag Agrawal, Omar Benjelloun, Anish Das Sarma, Chris Hayworth, Shubha Nabar, Tomoe Sugihara, and Jennifer Widom. Trio: A system for data, uncertainty, and lineage. In *Proceedings of the 32nd International Conference on Very Large Data Bases*, pages 1151–1154, Seoul, Korea, September 2006. (url: http://www.vldb.org/conf/2006/p1151-agrawal.pdf).

[2] Rocio Aldeco-Perez and Luc Moreau. Provenance-based auditing of private data use. In *BCS International Academic Research Conference, Visions of Computer Science*, September 2008. (url: http://eprints.ecs.soton.ac.uk/16580/).

[3] G. Alonso and A. El Abbadi. Goose: Geographic object oriented support environment. In *Proc. of the ACM workshop on Advances in Geographic Information Systems*, pages 38–49, Arlington, Virginia, November 1993.

[4] G. Alonso and C. Hagen. Geo-opera: Workflow concepts for spatial processes. In *Proceedings of 5th International Symposium on Spatial Databases (SSD'97)*, pages 238–258, Berlin, Germany, June 1997. (url: http://en.scientificcommons.org/216863).

[5] Ilkay Altintas, Oscar Barney, and Efrat Jaeger-Frank. Provenance collection support in the kepler scientific workflow system. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 118–132. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_14).

[6] Sergio Alvarez, Javier Vázquez-Salceda, Tamás Kifor, László Varga, and Steven Willmott. Applying provenance in distributed organ transplant management. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 28–36, 2006. (doi: http://dx.doi.org/10.1007/11890850_4).

[7] Manish Kumar Anand, Shawn Bowers, Timothy M. McPhillips, and Bertram Ludaescher. Efficient provenance storage over nested data collections. In Martin L. Kersten, Boris Novikov, Jens Teubner, Vladimir Polutin, and Stefan Manegold, editors, *Proceedings of teh 12th International Conference on Extending Database Technology (EDBT'09)*, pages 958–969, 2009. (doi: http://doi.acm.org/10.1145/1516360.1516470).

[8] Manish Kumar Anand, Shawn Bowers, Timothy M. McPhillips, and Bertram Ludaescher. Exploring scientific workflow provenance using hybrid queries over nested data and lineage graphs. In *Proceedings of 21st International Conference on Scientific and Statistical Database Management (SSDBM'09)*, pages 237–254, New Orleans, LA, USA, 2009. (doi: http://dx.doi.org/10.1007/978-3-642-02279-1_18).

[9] Erik W. Anderson, James P. Ahrens, Katrin Heitmann, Salman Habib, and Cl?udio T. Silva. Provenance in comparative analysis: A study in cosmology. *Computing in Science and Engineering*, 10(3):30–37, 2008. (doi: http://doi.ieeecomputersociety.org/10.1109/MCSE.2008.80).

[10] David W. Archer, Lois M. L. Delcambre, and David Maier. A framework for fine-grained data integration and curation, with provenance, in a dataspace. In James Cheney, editor, *TAPP'09: First workshop on on Theory and practice of provenance*, San Francisco, CA, February 2009. USENIX Association. (url: http://www.usenix.org/event/tapp09/tech/full_papers/archer/archer.pdf).

[11] Bartosz Balis, Marian Bubak, Michal Pelczar, and Jakub Wach. Provenance querying for end-users: A drug resistance case study. In *ICCS '08: Proceedings of the 8th international conference on Computational Science, Part III*, pages 80–89, Berlin, Heidelberg, 2008. Springer-Verlag. (doi: http://dx.doi.org/10.1007/978-3-540-69389-5_11).

[12] Bartosz Balis, Marian Bubak, Michal Pelczar, and Jakub Wach. Provenance tracking and querying in the virolab virtual laboratory. In *CCGRID '08: Proceedings of the 2008 Eighth IEEE International Symposium on Cluster Computing and the Grid*, pages 675–680, Washington, DC, USA, 2008. IEEE Computer Society. (doi: http://dx.doi.org/10.1109/CCGRID.2008.83).

[13] Bartosz Balis, Marian Bubak, and Jakub Wach. Provenance tracking in the virolab virtual laboratory. In *Parallel Processing and Applied Mathematics, 7th International Conference, PPAM 2007, Gdansk, Poland, September 9-12, 2007, Revised Selected Papers*, pages 381–390, 2007. (doi: http://dx.doi.org/10.1007/978-3-540-68111-3_40).

[14] Bartosz Balis, Marian Bubak, and Jakub Wach. User-oriented querying over repositories of data and provenance. In *E-SCIENCE '07: Proceedings of the Third IEEE International Conference on e-Science and Grid Computing*, pages 187–194, Washington, DC, USA, 2007. IEEE Computer Society. (doi: http://dx.doi.org/10.1109/E-SCIENCE.2007.81).

[15] Zhuowei Bao, Sarah Cohen-Boulakia, Susan B. Davidson, Anat Eyal, and Sanjeev Khanna. Differencing provenance in scientific workflows. In *IEEE 25th International Conference on Data Engineering (ICDE'09)*, pages 808–819. IEEE Computer Society, 2009. (doi: http://doi.ieeecomputersociety.org/10.1109/ICDE.2009.103).

[16] Roger S. Barga and Luciano A. Digiampietri. Automatic generation of workflow provenance. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 1–9. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_1).

[17] Roger S. Barga and Luciano A. Digiampietri. Automatic capture and efficient storage of escience experiment provenance. *Concurrency and Computation: Practice and Experience*, 20(5), 2008. (doi: http://dx.doi.org/10.1002/cpe.1235).

[18] Bruce R. Barkstrom. Data product configuration management and versioning in large-scale production of satellite scientific data production. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/CM_In_Lg_Scale_Production.doc).

[19] Bruce R. Barkstrom. Advances in provenance tracking and configuration management for earth science data. In *Eos Trans. American Geophysical Union, Fall Meeting 2008*, volume 89, 2008. abstract IN11C-1038, (url: http://www.agu.org/cgi-bin/wais?mm=IN11C-1038).

[20] Louis Bavoil, Steven P. Callahan, Patricia J. Crossno, Juliana Freire, Carlos E. Scheidegger, Claudio T. Silva, and Huy T. Vo. VisTrails: Enabling interactive multiple-view visualizations. In *In Proceedings of IEEE Visualization*, page 18, Los Alamitos, CA, USA, 2005. IEEE Computer Society. (doi: http://doi.ieeecomputersociety.org/10.1109/VIS.2005.113).

[21] R. A. Becker and J. M. J. M. Chambers. Auditing of data analyses. *SIAM Journal of Scientific and Statistical Computing*, 9(4):747–760, 1988. (doi: http://dx.doi.org/10.1137/0909049).

[22] Richard A. Becker and John M. Chambers. Auditing of data analyses. In Roger E. Cubitt, Brian Cooper, and Gultekin Ozsoyoglu, editors, *SSDBM'86: Proceedings of the 3rd international workshop on Statistical and*

scientific database management, pages 78–80, Berkeley, CA, US, 1986. Lawrence Berkeley Laboratory.

[23] Jeanne Behnke, John Moses, and James Byrnes. Archive issues associated with nasa earth science datasets. In *Eos Trans. American Geophysical Union, Fall Meeting 2008*, volume 89, 2008. abstract IN11C-1046, (url: http://www.agu.org/cgi-bin/wais?mm=IN11C-1046).

[24] Omar Benjelloun, Anish Das Sarma, Alon Halevy, Martin Theobald, and Jennifer Widom. Databases with uncertainty and lineage. *The VLDB Journal*, 17(2):243–264, 2008. (doi: http://dx.doi.org/10.1007/s00778-007-0080-z).

[25] Omar Benjelloun, Anish Das Sarma, Alon Halevy, and Jennifer Widom. Uldbs: databases with uncertainty and lineage. In *VLDB '06: Proceedings of the 32nd international conference on Very large data bases*, pages 953–964. VLDB Endowment, 2006. (url: http://ilpubs.stanford.edu:8090/703/).

[26] Omar Benjelloun, Anish Das Sarma, Chris Hayworth, and Jennifer Widom. An introduction to uldbs and the trio system. *IEEE Data Engineering Bulletin*, March 2006. (url: http://ilpubs.stanford.edu:8090/793/).

[27] Dave Berry, Peter Buneman, Michael Wilde, and Yannis Ioannidis, editors. *Data Provenance and Annotation*, Edinburgh, Scotland, December 2003. (url: http://www.nesc.ac.uk/esi/events/304/).

[28] Deepavali Bhagwat, Laura Chiticariu, Wang-Chiew Tan, and Gaurav Vijayvargiya. An annotation management system for relational databases. In *VLDB '04: Proceedings of the Thirtieth international conference on Very large data bases*, pages 900–911. VLDB Endowment, 2004. (url: http://www.vldb.org/conf/2004/RS23P1.PDF).

[29] Olivier Biton, Sarah Cohen Boulakia, and Susan B. Davidson. Zoom*userviews: Querying relevant provenance in workflow systems. In Christoph Koch, Johannes Gehrke, Minos N. Garofalakis, Divesh Srivastava, Karl Aberer, Anand Deshpande, Daniela Florescu, Chee Yong Chan, Venkatesh Ganti, Carl-Christian Kanne, Wolfgang Klas, and Erich J. Neuhold, editors, *VLDB '07: Proceedings of the 33rd international conference on Very large data bases*, pages 1366–1369. ACM, 2007. (url: http://www.vldb.org/conf/2007/papers/demo/p1366-biton.pdf).

[30] Olivier Biton, Sarah Cohen-Boulakia, Susan B. Davidson, and Carmem S. Hara. Querying and managing provenance through user views in scientific workflows. In *International Conference Data Engineering (ICDE'08)*, pages

1072–1081, Los Alamitos, CA, USA, 2008. IEEE Computer Society. (doi: http://doi.ieeecomputersociety.org/10.1109/ICDE.2008.4497516).

[31] Barbara T. Blaustein, Len Seligman, Michael Morse, M. David Allen, and Arnon Rosenthal. Plus: Synthesizing privacy, lineage, uncertainty and security. In *ICDE Workshops*, pages 242–245, 2008. (doi: http://dx.doi.org/10.1109/ICDEW.2008.4498325).

[32] Carsten Bochner, Roland Gude, and Andreas Schreiber. A python library for provenance recording and querying. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 229–240. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_24).

[33] R. Bose. A conceptual framework for composing and managing scientific data lineage. In *Proceedings of the 14th International Conference on Scientific and Statistical Database Management (SSDBM'02)*, pages 15–19, Washington, DC, USA, July 2002. IEEE Computer Society. (doi: http://dx.doi.org/10.1109/SSDM.2002.1029701).

[34] Raj Bose, Ian Foster, and Luc Moreau. Report on the international provenance and annotation workshop (ipaw'06). *Sigmod Records*, 35(3):51–53, September 2006. (doi: http://doi.acm.org/10.1145/1168092.1168102).

[35] Rajendra Bose and James Frew. Composing lineage metadata with xml for custom satellite-derived data products. In *16th International Conference on Scientific and Statistical Database Management*, pages 275–284, Santorini Island, Greece, June 2004. (doi: http://dx.doi.org/10.1109/SSDM.2004.1311219).

[36] Rajendra Bose and James Frew. Lineage retrieval for scientific data processing: A survey. *ACM Computing Surveys*, 37(1):1–28, March 2005. (doi: http://doi.acm.org/10.1145/1057977.1057978).

[37] Rajendra Bose, Robert G. Mann, and Diego Prina-Ricotti. Astrodas: Sharing assertions across astronomy catalogues through distributed annotation. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 193–202. Springer, May 2006. (doi: http://dx.doi.org/10.1007/11890850_20).

[38] Dimitri Bourilkov, Vaibhav Khandelwal, Archis Kulkarni, and Sanket Totala. Virtual logbooks and collaboration in science and software development. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume

4145 of *Lecture Notes in Computer Science*, pages 19–27. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_3).

[39] Shawn Bowers, Timothy McPhillips, Bertram Ludaescher, Shirley Cohen, and Susan B. Davidson. A model for user-oriented data provenance in pipelined scientific workflows. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 133–147. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_15).

[40] Shawn Bowers, Timothy McPhillips, Martin Wu, and Bertram Ludaescher. Project histories: Managing data provenance across collection-oriented scientific workflow runs. In *Proc. of the Intl. Workshop on Data Integration in the Life Sciences (DILS)*, volume 4544 of *Lecture Notes in Computer Science*, pages 122–138, 2007. (doi: http://dx.doi.org/10.1007/978-3-540-73255-6_12).

[41] Shawn Bowers, Timothy M. McPhillips, and Bertram Ludaescher. Provenance in collection-oriented scientific workflows. *Concurrency and Computation: Practice and Experience*, 20(5), 2008. (doi: http://dx.doi.org/10.1002/cpe.1226).

[42] Shawn Bowers, Timothy M. McPhillips, Sean Riddle, Manish Kumar Anand, and Bertram Ludaescher. Kepler/ppod: Scientific workflow and provenance support for assembling the tree of life. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 70–77. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_9).

[43] Miguel Branco and Luc Moreau. Enabling provenance on large scale e-science applications. In *Proceedings of the International Provenance and Annotation Workshop (IPAW'06)*, volume 4145 of *Lecture Notes in Computer Science*, pages 55–63, Chicago, Illinois, 2006. Springer-Verlag. (doi: http://dx.doi.org/10.1007/11890850_7).

[44] Uri Braun, Simson Garfinkel, David A. Holland, Kiran-Kumar Muniswamy-Reddy, and Margo I. Seltzer. Issues in automatic provenance collection. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 171–183. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_18).

[45] Uri Braun, David A. Holland, Kiran-Kumar Muniswamy-Reddy, and Margo I. Seltzer. Coping with cycles in provenance. Technical report, Harvard University, 2006. (url: http://www.eecs.harvard.edu/~syrah/pubs/cycles.pdf).

[46] Uri Braun and Avi Shinnar. A security model for provenance. Technical Report TR-04-06, Harvard University Computer Science, January 2006. (url: ftp://ftp.deas.harvard.edu/techreports/tr-04-06.pdf).

[47] Uri Braun, Avraham Shinnar, and Margo Seltzer. Securing provenance. In *HOTSEC'08: Proceedings of the 3rd conference on Hot topics in security*, pages 1–5, Berkeley, CA, USA, 2008. USENIX Association. (url: http://www.usenix.org/event/hotsec08/tech/full_papers/braun/braun.pdf).

[48] Allen L. Brown. Enforcing the scientific method. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, page 2. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_2).

[49] P. Buneman, S. Khanna, K.Tajima, and W.C. Tan. Archiving scientific data. In *Proc. of the 2002 ACM SIGMOD International Conference on Management of Data*, pages 1–12. ACM Press, 2002. (doi: http://doi.acm.org/10.1145/564691.564693).

[50] P. Buneman, D. Maier, and J. Widom. Where was your data yesterday, and where will it go tomorrow? In *Position paper for NSF Workshop on Information and Data Management (IDM '00), Chicago IL*, 2000. (url: http://hermes.dpi.inpe.br:1910/col/dpi.inpe.br/banon/2004/04.21.11.45/doc/BunemanWhereTomorrow.pdf).

[51] Peter Buneman, Adriane Chapman, and James Cheney. Provenance management in curated databases. In *SIGMOD '06: Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 539–550, New York, NY, USA, 2006. ACM Press. (doi: http://doi.acm.org/10.1145/1142473.1142534).

[52] Peter Buneman, Adriane Chapman, James Cheney, and Stijn Vansummeren. A provenance model for manually curated data. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 162–170. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_17).

[53] Peter Buneman, James Cheney, Wang-Chiew Tan, and Stijn Vansummeren. Curated databases. In *PODS '08: Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 1–12, New York, NY, USA, 2008. ACM. (doi: http://doi.acm.org/10.1145/1376916.1376918).

[54] Peter Buneman, James Cheney, and Stijn Vansummeren. On the expressiveness of implicit provenance in query and update languages. In *11th International Conference on Database Theory (ICDT 2007)*, volume 4353 of *Lecture Notes in Computer Science*, pages 209–223, 2007. (doi: http://dx.doi.org/10.1007/11965893_15).

[55] Peter Buneman, James Cheney, and Stijn Vansummeren. On the expressiveness of implicit provenance in query and update languages. *ACM Trans. Database Syst.*, 33(4):1–47, 2008. (doi: http://doi.acm.org/10.1145/1412331.1412340).

[56] Peter Buneman, Sanjeev Khanna, Keishi Tajima, and Wang-Chiew Tan. Data archiving. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/pp.ps).

[57] Peter Buneman, Sanjeev Khanna, Keishi Tajima, and Wang Chiew Tan. Archiving scientific data. *ACM Trans. Database Syst.*, 29:2–42, 2004. (doi: http://doi.acm.org/10.1145/974750.974752).

[58] Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Data provenance: Some basic issues. In *FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science*, volume 1974 of *Lecture Notes in Computer Science*, pages 87–93, 2000. (doi: http://dx.doi.org/10.1007/3-540-44450-5).

[59] Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Why and Where: A Characterization of Data Provenance. In *Proceedings of 8th International Conference on Database Theory (ICDT'01)*, volume 1973 of *Lecture Notes in Computer Science*, pages 316–330, London, UK, 2001. Springer. (doi: http://dx.doi.org/10.1007/3-540-44503-X_20).

[60] Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Computing provenance and annotations for views. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/provenance_s.ps).

[61] Peter Buneman and Dan Suciu. Letter from the special issue editor. *IEEE Data Eng. Bull.*, 30(4):2, 2007. (url: http://sites.computer.org/debull/A07dec/letter-peter-dan.pdf).

[62] Peter Buneman and Wang-Chiew Tan. Provenance in databases. In *SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 1171–1173, New York, NY, USA, 2007. ACM. (doi: http://doi.acm.org/10.1145/1247480.1247646).

[63] Steven P. Callahan, Juliana Freire, Juliana Freire, Emanuele Santos, Carlos E. Scheidegger, Claudio T. Silva, and Huy T. Vo. Managing the evolution of dataflows with vistrails. In *Data Engineering Workshops, 22nd International Conference on*, page 71, Los Alamitos, CA, USA, 2006. IEEE Computer Society. (doi: http://doi.ieeecomputersociety.org/10.1109/ICDEW.2006.75).

[64] Steven P. Callahan, Juliana Freire, Emanuele Santos, Carlos E. Scheidegger, Claudio T. Silva, and Huy T. Vo. Using provenance to streamline data exploration through visualization. Technical Report USCI-2006-17, University of Utah, 2006. (url: http://www.sci.utah.edu/publications/SCITechReports/UUSCI-2006-016.pdf).

[65] Steven P. Callahan, Juliana Freire, Carlos Eduardo Scheidegger, Cláudio T. Silva, and Huy T. Vo. Towards provenance-enabling paraview. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 120–127. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_13).

[66] Jeremy J. Carroll, Christian Bizer, Pat Hayes, and Patrick Stickler. Named graphs, provenance and trust. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 613–622, New York, NY, USA, 2005. ACM Press. (doi: http://doi.acm.org/10.1145/1060745.1060835).

[67] Maria Claudia Cavalcanti, Maria Luiza Campos, and Marta Mattoso. Managing scientific models in structural genomic projects updated. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/CCM.ps).

[68] Richard Cavanaugh and Greg Graham. Apples and apple-shaped oranges: Equivalence of data returned on subsequent queries with provenance information. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/apples-oranges.ps).

[69] Richard Cavanaugh, Greg Graham, and Mike Wilde. Satisfying the tax collector: Using data provenance as a way to audit data analyses in high energy physics. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/TAXMan.ps).

[70] http://twiki.ipaw.info/bin/view/Challenge/FirstProvenanceChallenge, June 2006.

[71] The provenance challenge wiki. http://twiki.ipaw.info/bin/view/Challenge, June 2006.

[72] Adriane Chapman. *Incorporating Provenance in Database Systems*. PhD thesis, University of Michigan, 2008. (url: http://hdl.handle.net/2027.42/61645).

[73] Adriane Chapman and H. V. Jagadish. Issues in building practical provenance systems. *IEEE Data Eng. Bull.*, 30(4):38–43, 2007. (url: http://sites.computer.org/debull/A07dec/chapman.pdf).

[74] Adriane Chapman and H. V. Jagadish. Provenance and the price of identity. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 106–119. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_12).

[75] Adriane P. Chapman, H. V. Jagadish, and Prakash Ramanan. Efficient provenance storage. In *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 993–1006, New York, NY, USA, 2008. ACM. (doi: http://doi.acm.org/10.1145/1376616.1376715).

[76] A. Chebotko, S. Chang, S. Lu, F. Fotouhi, and P. Yang. Scientific workflow provenance querying with security views. In *In Proceedings of the Ninth International Conference on Web-Age Information Management*, pages 349–356. IEEE Computer Society, 2008. (doi: http://dx.doi.org/10.1109/WAIM.2008.41).

[77] Artem Chebotko, Xubo Fei, Cui Lin, Shiyong Lu, and Farshad Fotouhi. Storing and querying scientific workflow provenance metadata using an rdbms. In *e-Science and Grid Computing, International Conference on*, volume 0, pages 611–618, Los Alamitos, CA, USA, 2007. IEEE Computer Society. (doi: http://doi.ieeecomputersociety.org/10.1109/E-SCIENCE.2007.70).

[78] Liming Chen and Zhuoan Jiao. Supporting provenance in service-oriented computing using the semantic web technologies. *IEEE Intelligent Informatics Bulletin*, 7(1):4–11, 2006. (url: http://www.comp.hkbu.edu.hk/~cib/2006/Dec/iib_vol7no1_article1.pdf).

[79] Liming Chen, Zhuoan Jiao, and Simon J. Cox. On the use of semantic annotations for supporting provenance in grids. Springer, 2006. (doi: http://dx.doi.org/10.1007/11823285).

[80] Liming Chen, Victor Tan, Fenglian Xu, Alexis Biller, Paul Groth, Simon Miles, John Ibbotson, Michael Luck, and Luc Moreau. A proof of concept: Provenance in a service oriented architecture. In *Proceedings of the Fourth All Hands Meeting (AHM)*, September 2005. (url: http://www.allhands.org.uk/2005/proceedings/papers/503.pdf).

[81] Liming Chen, Xueqiang Yang, and Feng Tao. A semantic web service based approach for augmented provenance. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 594–600, Washington, DC, USA, 2006. IEEE Computer Society. (doi: http://dx.doi.org/10.1109/WI.2006.25).

[82] Zheng Chen and Luc Moreau. Implementation and evaluation of a protocol for recording process documentation in the presence of failures. In *Proceedings of Second International Provenance and Annotation Workshop (IPAW'08)*, volume 5272 of *Lecture Notes in Computer Science*, pages 92–105, Salt Lake City, USA, June 2008. Springer-Verlag. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_11).

[83] Zheng Chen and Luc Moreau. Recording process documentation in the presence of failures. In *Methods, Models and Tools for Fault Tolerance*, volume 5454 of *Lecture Notes in Computer Science*, pages 196–219. Springer-Verlag, 2009. (doi: http://dx.doi.org/10.1007/978-3-642-00867-2_10).

[84] J. Cheney, A. Ahmed, and U. Acar. Provenance as dependency analysis. *Under consideration for publication in Math. Struct. in Comp. Science*, March 2008. (url: http://homepages.inf.ed.ac.uk/jcheney/publications/drafts/prov-dep-jv.pdf).

[85] James Cheney. Program slicing and data provenance. *IEEE Data Eng. Bull.*, 30(4):22–28, 2007. (url: http://sites.computer.org/debull/A07dec/cheney.pdf).

[86] James Cheney, editor. *First Workshop on the Theory and Practice of Provenance, 2009, San Francisco, CA*, San Francisco, CA, February 2009. USENIX Association. (url: http://www.usenix.org/events/tapp09/).

[87] James Cheney. Provenance, xml, and the scientific web. In *Programming Language Techniques for XML (Plan-X'09)*, 2009. (url: http://db.ucsd.edu/planx2009/camera-ready/unpaginated/invited.pdf).

[88] James Cheney, Umut A. Acar, and Amal Ahmed. Provenance traces (extended report). Technical Report http://arxiv.org/abs/0812.0564v1, University of Edinburgh, December 2008. (url: http://homepages.inf.ed.ac.uk/jcheney/publications/drafts/provenance-traces-tr.pdf).

[89] James Cheney, Amal Ahmed, and Umut A. Acar. Provenance as dependency analysis. In M. Arenas and M. I. Schwartzbach, editors, *Proceedings of the 11th International Symposium on Database Programming Languages (DBPL 2007)*, number 4797 in Lecture Notes in Computer Science, pages 139–153, 2007. (doi: 10.1007/978-3-540-75987-4_10).

[90] James Cheney, Peter Buneman, and Bertram Ludaescher. Report on the principles of provenance workshop. *SIGMOD Record*, 37(1):62–65, 2008. (doi: http://doi.acm.org/10.1145/1374780.1374798).

[91] K Cheung and J Hunter. Provenance explorer - customized provenance views using semantic inferencing. In *5th International Semantic Web Conference (ISWC2006)*, volume 4273 of *Lecture Notes in Computer Science*. Springer-Verlag, 2006. (doi: http://dx.doi.org/10.1007/11926078_16).

[92] Laura Chiticariu and Wang-Chiew Tan. Debugging schema mappings with routes. In *VLDB '06: Proceedings of the 32nd international conference on Very large data bases*, pages 79–90. VLDB Endowment, 2006. (url: http://www.vldb.org/conf/2006/p79-chiticariu.pdf).

[93] Laura Chiticariu, Wang-Chiew Tan, and Gaurav Vijayvargiya. Dbnotes: a post-it system for relational databases based on provenance. In *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 942–944, New York, NY, USA, 2005. ACM. (doi: http://doi.acm.org/10.1145/1066157.1066296).

[94] Stephen Chong. Towards semantics for provenance security. In James Cheney, editor, *TAPP'09: First workshop on on Theory and practice of provenance*, San Francisco, CA, February 2009. USENIX Association. (url: http://www.usenix.org/event/tapp09/tech/full_papers/chong/chong.pdf).

[95] Alison Chorley, Pete Edwards, Alun Preece, and John Farrington. Tools for tracing evidence in social science. In *Third International Conference on e-Social Science*, October 2007. (url: http://users.cs.cf.ac.uk/A.D.Preece/publications/download/ess2007b.pdf).

[96] Andrew Cirillo, Radha Jagadeesan, Corin Pitcher, and James Riely. Tapido: Trust and authorization via provenance and integrity in distributed objects (extended abstract). In Sophia Drossopoulou, editor, *7th European Symposium on Programming (ESOP'08)*, volume 4960 of *Lecture*

Notes in Computer Science, pages 208–223. Springer, 2008. (doi: http://dx.doi.org/10.1007/978-3-540-78739-6_17).

[97] Ben Clifford, Ian Foster, Mihael Hategan, Tiberiu Stef-Praun, Michael Wilde, and Yong Zhao. Tracking provenance in a virtual data grid. *Concurrency and Computation: Practice and Experience*, 20(5):565–575, 2008. (doi: http://dx.doi.org/10.1002/cpe.1256).

[98] Shirley Cohen, Sarah Cohen-Bolakia, and Susan B. Davidson. Towards a model of provenance and user views in scientific workflows. In *Third International Workshop on Data Integration in the Life Sciences (DIL'06)*, volume 4076 of *Lecture Notes in Computer Science*, pages 264–279, Hinxton, UK, July 2006. Springer. (doi: http://dx.doi.org/10.1007/11799511).

[99] Sarah Cohen-Boulakia, Olivier Biton, Shirley Cohen, and Susan Davidson. Addressing the provenance challenge using zoom. *Concurrency and Computation: Practice and Experience*, 20(5):497–506, 2008. (doi: http://dx.doi.org/10.1002/cpe.1232).

[100] Simon Cox, Rachel Jones, Bryan Lawrence, Natasa Milic-Frayling, and Luc Moreau. Interoperability issues in scientific data management (version 1.0). Technical report, The Technical Computing Initiative, Microsoft Corporation, March 2006. (url: http://science.officeisp.net/SharedDocuments/ScientificDataManagement4.18.07.pdf).

[101] Daniel Crawl and Ilkay Altintas. A provenance-based fault tolerance mechanism for scientific workflows. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 152–159. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_17).

[102] Y. Cui. *Lineage Tracing in Data Warehouses*. PhD thesis, Stanford University, December 2001. (url: http://ilpubs.stanford.edu:8090/522/).

[103] Y. Cui and J. Widom. Lineage tracing in a data warehousing system. In *Proceedings of the 16th International Conference on Data Engineering*, pages 683–684, San Diego, California, 2000. Demonstration Description, (doi: http://dx.doi.org/10.1109/ICDE.2000.839493).

[104] Y. Cui and J. Widom. Practical lineage tracing in data warehouses. In *Proceedings of the 16th International Conference on Data Engineering (ICDE'00)*, pages 367–378, San Diego, California, February 2000. (doi: http://dx.doi.org/10.1109/ICDE.2000.839437).

[105] Y. Cui and J. Widom. Storing auxiliary data for efficient maintenance and lineage tracing of complex views. In *Proceedings of the International Workshop on Design and Management of DataWarehouses (DMDW'00)*, Stockholm, Sweden, 2000. (url: http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-28/paper11.pdf).

[106] Y. Cui and J. Widom. Lineage tracing for general data warehouse transformations. *The VLDB Journal*, 12(1):41–58, 2003. (doi: http://dx.doi.org/10.1007/s00778-002-0083-8).

[107] Y. Cui, J. Widom, and J. L. Wiener. Tracing the lineage of view data in a warehousing environment. *ACM Trans. Database Syst.*, 25(2):179–227, 2000. (doi: http://doi.acm.org/10.1145/357775.357777).

[108] Francisco Curbera, Yurdaer N. Doganata, Axel Martens, Nirmal Mukhi, and Aleksander Slominski. Business provenance - a technology to increase traceability of end-to-end operations. In *On the Move to Meaningful Internet Systems: OTM'2008*, pages 100–119, 2008. (doi: http://dx.doi.org/10.1007/978-3-540-88871-0_10).

[109] Bryce Cutt and Ramon Lawrence. Managing data quality in a terabyte-scale sensor archive. In *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, pages 982–986, New York, NY, USA, 2008. ACM. (doi: http://doi.acm.org/10.1145/1363686.1363915).

[110] Sérgio Manuel Serra da Cruz, Fernando Seabra Chirigati, Rafael Dahis, Maria Luiza Machado Campos, and Marta Mattoso. Using explicit control processes in distributed workflows to gather provenance. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 186–199. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_20).

[111] S.M.S. da Cruz, P.M. Barros, P.M. Bisch, M.L.M. Campos, and M. Mattoso. Provenance services for distributed workflows. In *Cluster Computing and the Grid, 2008. CCGRID '08. 8th IEEE International Symposium on*, pages 526–533, May 2008. (doi: http://dx.doi.org/10.1109/CCGRID.2008.73).

[112] Srgio Manuel Serra da Cruz, Maria Luiza M. Campos, and Marta Mattoso. Towards a taxonomy of provenance in scientific workflow management systems. volume 0, pages 259–266, Los Alamitos, CA, USA, 2009. IEEE Computer Society. (doi: http://doi.ieeecomputersociety.org/10.1109/SERVICES-I.2009.18).

[113] Paulo Pinheiro da Silva, Deborah L. McGuinness, and Rob McCool. Knowledge provenance infrastructure. *IEEE Data Engineering Bulletin*, 26(4):26–32, December 2003. (url: http://www-ksl.stanford.edu/people/pp/papers/PinheirodaSilva_DEBULL_2003.pdf).

[114] Chenyun Dai, Dan Lin, Elisa Bertino, and Murat Kantarcioglu. An approach to evaluate data trustworthiness based on data provenance. In *SDM '08: Proceedings of the 5th VLDB workshop on Secure Data Management*, pages 82–98, Berlin, Heidelberg, 2008. Springer-Verlag. (doi: http://dx.doi.org/10.1007/978-3-540-85259-9_6).

[115] Stephen Davey, Ali Anjomshoaa, Mario Antonioletti, Malcolm Atkinson, Dave Berry, Ann Chervenak, Adrian Jackson, Chris Jordan, Peter Kunszt, Allen Luniewski, and Luc Moreau. Ogsa data scenarios v0.13. Technical report, Global Grid Forum, June 2006. (url: https://forge.gridforum.org/sf/go/doc13605?nav=1).

[116] Susan B. Davidson, Sarah Cohen Boulakia, Anat Eyal, Bertram Ludaescher, Timothy M. McPhillips, Shawn Bowers, Manish Kumar Anand, and Juliana Freire. Provenance in scientific workflow systems. *IEEE Data Eng. Bull.*, 30(4):44–50, 2007. (url: http://sites.computer.org/debull/A07dec/susan.pdf).

[117] Susan B. Davidson and Juliana Freire. Provenance and scientific workflows: challenges and opportunities. In *SIGMOD Conference*, pages 1345–1350, 2008. (doi: http://doi.acm.org/10.1145/1376616.1376772).

[118] Frederico T. de Oliveira, Leonardo Gresta Paulino Murta, Cláudia Werner, and Marta Mattoso. Using provenance to improve workflow design. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 136–143. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_15).

[119] Ewa Deelman, Scott Callaghan, Edward Field, Hunter Francoeur, Robert Graves, Nitin Gupta, Vipin Gupta, Thomas H. Jordan, Carl Kesselman, Philip Maechling, John Mehringer, Gaurang Mehta, David Okaya, Karan Vahi, and Li Zhao. Managing large-scale workflow execution from resource provisioning to provenance tracking: The cybershake example. In *E-SCIENCE '06: Proceedings of the Second IEEE International Conference on e-Science and Grid Computing*, page 14, Washington, DC, USA, 2006. IEEE Computer Society. (doi: http://dx.doi.org/10.1109/E-SCIENCE.2006.99).

[120] Vinay Deolalikar and Hernan Laffitte. Provenance as data mining: combining file system metadata with content analysis. In James Cheney, editor, *TAPP'09: First workshop on on Theory and practice of provenance*, San Francisco, CA, February 2009. USENIX Association. (url: http://www.usenix.org/event/tapp09/tech/full_papers/deolalikar/deolalikar.pdf).

[121] Vikas Deora, Arnaud Contes, Omer F. Rana, Shrija Rajbhandari, Ian Wootten, Kifor Tamas, and Laszlo Z.Varga. Navigating provenance information for distributed healthcare management. In *IEEE/WIC/ACM Web Intelligence Conference*, pages 859–865, Washington, DC, USA, 2006. IEEE Computer Society. (doi: http://dx.doi.org/10.1109/WI.2006.122).

[122] L Di and Peng Yue. Geospatial data provenance in the semantic web environment. In *Eos Trans. American Geophysical Union, Fall Meeting 2008*, volume 89, 2008. abstract IN11C-1043, (url: http://www.agu.org/cgi-bin/wais?mm=IN11C-1043).

[123] Li Ding, Tim Finin, Yun Peng, Paulo Pinheiro da Silva, and Deborah L. McGuinness. Tracking RDF Graph Provenance using RDF Molecules. Technical report, UMBC, April 2005. (url: ftp://ksl.stanford.edu/pub/KSL_Reports/KSL-05-06.pdf).

[124] Li Ding, Pranam Kolari, Tim Finin, Anupam Joshi, Yun Peng, and Yelena Yesha. On homeland security and the semantic web: A provenance and trust aware inference framework. In *Proceedings of the AAAI SPring Symposium on AI Technologies for Homeland Security*. AAAI Press, 2005. (url: http://ebiquity.umbc.edu/paper/html/id/209/On-Homeland-Security-and-the-Semantic-Web-A-Provenance-and-Trust-Aware-Inference

[125] Renata Dividino, Sergej Sizov, Steffen Staab, and Bernhard Schueler. Querying for provenance, trust, uncertainty and other meta knowledge in rdf. *Web Semantics: Science, Services and Agents on the World Wide Web*, In Press, Corrected Proof:–, 2009. (doi: http://dx.doi.org/10.1016/j.websem.2009.07.004).

[126] Renata Queiroz Dividino, Simon Schenk, Sergej Sizov, and Steffen Staab. Provenance, trust, explanations - and all that other meta knowledge. *KI*, 23(2):24–30, 2009. (url: http://www.uni-koblenz.de/~staab/Research/Publications/2009/provenance-ki-2009.pdf).

[127] Andrew Dolgert, Lawrence Gibbons, Christopher D. Jones, Valentin Kuznetsov, Mirek Riedewald, Daniel Riley, Gregory J. Sharp, and Peter Wittich. Provenance in high-energy physics workflows. *Computing in Science and Engineering*, 10(3):22–29, 2008. (doi: http://doi.ieeecomputersociety.org/10.1109/MCSE.2008.81).

[128] Jeff Dozier and James Frew. Computational provenance in hydrologic science: a snow mapping example. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1890):1021–1033, 2009. (doi: http://dx.doi.org/10.1098/rsta.2008.0187).

[129] James R. Driscoll, Neil Sarnak, Daniel D. Sleator, and Robert E. Tarjan. Making data structures persistent. *Journal of Computer and System Sciences*, 38(1):86 – 124, 1989. (doi: http://dx.doi.org/10.1016/0022-0000(89)90034-2).

[130] Ruth E. Duerr. Provenance: Promise and practice. In *Eos Trans. American Geophysical Union, Fall Meeting 2008*, volume 89, 2008. abstract IN11C-1040, (url: http://www.agu.org/cgi-bin/wais?mm=IN11C-1040).

[131] Frantisek Dvorak, Daniel Kouril, Ales Krenek, Ludek Matyska, Milos Mulac, Jan Pospisil, Miroslav Ruda, Zdenek Salvet, Jiri Sitera, and Michal Vocu. glite job provenance. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 246–253. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_25).

[132] P. D. Eagan and Ventura. Enhancing value of environmental data: data lineage reporting. *Journal of Environmental Engineering*, 119(1):5–16, 2007. (doi: 10.1061/(ASCE)0733-9372(1993)119:1(5)).

[133] Johann Eder, Georg E. Olivotto, and Wolfgang Gruber. A data warehouse for workflow logs. In Y.Han, S.Tai, and D.Wikarski, editors, *Engineering and Deployment of Cooperative Information Systems: First Int. Conf., EDCIS 2002*, volume 2480 of *Lecture Notes in Computer Science*. Springer, September 2002. (doi: http://dx.doi.org/10.1007/3-540-45785-2_1).

[134] Tommy Ellkvist, David Koop, Erik W. Anderson, Juliana Freire, and Cláudio T. Silva. Using provenance to support real-time collaborative design of workflows. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 266–279. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_27).

[135] Michael Factor, Ealan Henis, Dalit Naor, Simona Rabinovici-Cohen, Petra Reshef, Shahar Ronen, Giovanni Michetti, and Maria Guercio. Authenticity and provenance in long term digital preservation: modeling and implementation in preservation aware storage. In James Cheney, editor, *TAPP'09: First workshop on on Theory and practice of provenance*, San Francisco,

CA, February 2009. USENIX Association. (url: http://www.usenix.org/event/tapp09/tech/full_papers/factor/factor.pdf).

[136] Hao Fan. Tracing data lineage using automed schema transformation pathways. In *Advances in Databases*, pages 44–55. Springer-Verlag, 2002. (doi: http://dx.doi.org/10.1007/3-540-45495-0_6).

[137] Hao Fan and Alexandra Poulovassilis. Tracing data lineage using schema transformation pathways. In B. Omelayenko and M. Klein, editors, *Knowledge Transformation for the Semantic Web. Frontiers in Artificial Intelligence and Applications*, pages 64–79. IOS Press, 2003. (url: http://www.doc.ic.ac.uk/automed/publications/FP03a.ps).

[138] Hao Fan and Alexandra Poulovassilis. Using schema transformation pathways for data lineage tracing. In *Database: Enterprise, Skills and Innovation*, volume 3567, pages 133–144, June 2005. (doi: http://dx.doi.org/10.1007/11511854_11).

[139] Yuhong Feng and Wentong Cai. Provenance provisioning in mobile agent-based distributed job workflow execution. In *ICCS '07: Proceedings of the 7th international conference on Computational Science, Part I*, volume 4487 of *Lecture Notes in Computer Science*, pages 398–405, Berlin, Heidelberg, 2007. Springer-Verlag. (doi: http://dx.doi.org/10.1007/978-3-540-72584-8_51).

[140] Open provenance model workshop: Towards provenance challenge 3. http://twiki.ipaw.info/bin/view/Challenge/OpenProvenanceModelWorkshop, June 2008.

[141] Albert J. Fleig. Current climate data set documentation standards: Somewhere between anagrams and full disclosure. In *Eos Trans. American Geophysical Union, Fall Meeting 2008*, volume 89, 2008. abstract IN11C-1045, (url: http://www.agu.org/cgi-bin/wais?mm=IN11C-1045).

[142] Albert J. Fleig. Source code, an essential part of providing complete provenance. In *Eos Trans. American Geophysical Union, Fall Meeting 2008*, volume 89, 2008. abstract IN11C-1044, (url: http://www.agu.org/cgi-bin/wais?mm=IN11C-1044).

[143] The R Foundation for Statistical Computing. R: Regulatory compliance and validation issues a guidance document for the use of r in regulated clinical trial environments. Technical report, Wirtschaftsuniversiat Wien, 2008. (url: http://www.r-project.org/doc/R-FDA.pdf).

[144] Ian Foster and Peter Buneman. Workshop on data provenance and derivation. October 2002, (url: http://people.cs.uchicago.edu/~yongzh/position_papers.html).

[145] Ian Foster, Jens Vockler, Michael Wilde, and Yong Zhao. The virtual data grid: A new model and architecture for data-intensive collaboration. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/CIDR.VDG.submitted.pdf).

[146] Ian Foster, Jens-S. Voeckler, Michael Wilde, and Yong Zhao. Chimera: A virtual data system for representing, querying and automating data derivation. In *Proceedings of the 14th Conference on Scientific and Statistical Database Management (SSDBM'02)*, pages 37–46, Edinburgh, Scotland, July 2002. (doi: http://doi.ieeecomputersociety.org/10.1109/SSDM.2002.1029704).

[147] J. Nathan Foster, Todd J. Green, and Val Tannen. Annotated xml: queries and provenance. In *PODS '08: Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 271–280, New York, NY, USA, 2008. ACM. (doi: http://doi.acm.org/10.1145/1376916.1376954).

[148] J. Nathan Foster and Grigoris Karvounarakis. Provenance and data synchronization. *IEEE Data Eng. Bull.*, 30(4):13–21, 2007. (url: http://sites.computer.org/debull/A07dec/foster.pdf).

[149] Geoffrey Fox and David Walker. e-science gap analysis. Technical report, National e-Science Center, 2003. (url: http://www.nesc.ac.uk/technical_papers/UKeS-2003-01/index.html).

[150] Mark S. Fox and Jingwei Huang. Knowledge provenance. In *Advances in Artificial Intelligence. 17th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2004, London, Ontario, Canada, May 17-19, 2004*, volume 3060 of *Lecture Notes in Computer Science*. Springer, 2004. (doi: http://dx.doi.org/10.1007/b97823).

[151] Mark S. Fox and Jingwei Huang. An ontology for static knowledge provenance. In *Knowledge Sharing in the Integrated Enterprise*, IFIP International Federation for Information Processing, pages 203–213, 2005. (doi: http://dx.doi.org/10.1007/0-387-29766-9_17).

[152] M.S. Fox and J. Huang. Knowledge provenance in enterprise information. *International Journal of Production Research*, 43(20):4471–4492, October 2005. (doi: http://dx.doi.org/10.1080/00207540500142415).

[153] Peter Fox. Some thoughts on data derivation and provenance. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/provenance.ps).

[154] Juliana Freire, David Koop, and Luc Moreau, editors. *Provenance and Annotation of Data — International Provenance and Annotation Workshop, IPAW 2008*, volume 5272 of *Lecture Notes in Computer Science*. Springer-Verlag, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5).

[155] Juliana Freire, David Koop, Emanuele Santos, and Claudio T. Silva. Provenance for computational tasks: A survey. *Computing in Science and Engineering*, 10(3):11–21, 2008. (doi: http://doi.ieeecomputersociety.org/10.1109/MCSE.2008.79).

[156] Juliana Freire, Claudio T. Silva, Steven P. Callahan, Emanuele Santos, Carlos E. Scheidegger, and Huy T. Vo. Managing rapidly-evolving scientific workflows. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 10–18. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_2).

[157] J. Frew and R. Bose. Earth system science workbench: A data management infrastructure for earth science products. In *Proceedings of the 13th International Conference on Scientific and Statistical Database Management (SSDBM'01)*, pages 180–189, Fairfax, VA, July 2001. (doi: http://dx.doi.org/10.1109/SSDM.2001.938550).

[158] James Frew and Rajendra Bose. Lineage issues for scientific data and information. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/position-paper.html).

[159] James Frew, Dominic Metzger, and Peter Slaughter. Automatic capture and reconstruction of computational provenance. *Concurrency and Computation: Practice and Experience*, 20(5):485–496, 2008. (doi: http://dx.doi.org/10.1002/cpe.1247).

[160] James Frew and Peter Slaughter. Automatic run-time provenance capture for scientific dataset generation. In *Eos Trans. American Geophysical Union, Fall Meeting 2008*, volume 89, 2008. abstract IN11C-1039, (url: http://www.agu.org/cgi-bin/wais?mm=IN11C-1039).

[161] James Frew and Peter Slaughter. Es3: A demonstration of transparent provenance for scientific computation. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer*

Science, pages 200–207. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_21).

[162] Jeremy Frey, David De Roure, Kieron Taylor, Jonathan Essex, Hugo Mills, and Ed Zaluska. Combechem: A case study in provenance and annotation using the semantic web. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 270–277. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_27).

[163] Joe Futrelle. Harvesting rdf triples. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 64–72. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_8).

[164] Joe Futrelle. Tupelo semantic content repository – tutorial on provenance. http://tupeloproject.ncsa.uiuc.edu/node/2, 2008.

[165] Joe Futrelle and Jim Myers. Tracking provenance semantics in heterogeneous execution systems. *Concurrency and Computation: Practice and Experience*, 20(5):555–564, 2008. (doi: http://dx.doi.org/10.1002/cpe.1253).

[166] Luiz M. R. Gadelha and Marta Mattoso. Kairos: An architecture for securing authorship and temporal information of provenance data in grid enabled workflow management systems. In *e-Science and Grid Computing, International Conference on*, volume 0, pages 597–602, Los Alamitos, CA, USA, 2008. IEEE Computer Society. (doi: http://doi.ieeecomputersociety.org/10.1109/eScience.2008.161).

[167] Helena Galhardas, Daniela Florescu, Dennis Shasha, Eric Simon, and Cristian augustin Saita. Improving data cleaning quality using a data lineage facility. In *In: Proc. Design and Management of Data Warehouses (DMDW'01)*, 2001. (url: http://citeseer.ist.psu.edu/451787.html).

[168] A. Gehani and U. Lindqvist. Bonsai: Balanced lineage authentication. In *Computer Security Applications Conference, 2007. ACSAC 2007. Twenty-Third Annual*, pages 363–373, Dec. 2007. (doi: http://dx.doi.org/10.1109/ACSAC.2007.45).

[169] Ashish Gehani, Minyoung Kim, and Jian Zhang. Steps toward managing lineage metadata in grid clusters. In James Cheney, editor, *TAPP'09: First workshop on on Theory and practice of provenance*, San Francisco,

CA, February 2009. USENIX Association. (url: http://www.usenix.org/event/tapp09/tech/full_papers/gehani/gehani.pdf).

[170] Ashish Gehani and Ulf Lindqvist. Veil: A system for certifying video provenance. In *ISM '07: Proceedings of the Ninth IEEE International Symposium on Multimedia*, pages 263–272, Washington, DC, USA, 2007. IEEE Computer Society. (doi: http://dx.doi.org/10.1109/ISM.2007.10).

[171] Michael Gertz. Data annotations in collaborative research environments. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/gertz-position-paper.pdf).

[172] Tara Gibson, Karen Schuchardt, and Eric Stephan. Application of named graphs towards custom provenance views. In James Cheney, editor, *TAPP'09: First workshop on on Theory and practice of provenance*, San Francisco, CA, February 2009. USENIX Association. (url: http://www.usenix.org/event/tapp09/tech/full_papers/gibson/gibson.pdf).

[173] Tara Gibson, Karen Schuchardt, and Eric G. Stephan. Application of provenance for automated and research driven workflows. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 128–135. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_14).

[174] Yolanda Gil, Ewa Deelman, Mark Ellisman, Thomas Fahringer, Geoffrey Fox, Dennis Gannon, Carole Goble, Miron Livny, Luc Moreau, and Jim Myers. Examining the challenges of scientific workflows. *IEEE Computer*, 40(12):26–34, December 2007. (doi: http://doi.ieeecomputersociety.org/10.1109/MC.2007.421).

[175] Yolanda Gil, Varun Ratnakar, and Ewa Deelman. Metadata catalogs with semantic representations. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 90–100. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_11).

[176] Boris Glavic and Gustavo Alonso. Perm: Processing provenance and data on the same data model through query rewriting. In *IEEE 25th International Conference on Data Engineering (ICDE'09)*, pages 174–185. IEEE Computer Society, 2009. (doi: http://doi.ieeecomputersociety.org/10.1109/ICDE.2009.15).

[177] Boris Glavic and Gustavo Alonso. Provenance for nested subqueries. In *EDBT '09: Proceedings of the 12th International Conference on Extending*

Database Technology, pages 982–993, New York, NY, USA, 2009. ACM. (doi: http://doi.acm.org/10.1145/1516360.1516472).

[178] Boris Glavic and Klaus R. Dittrich. Data provenance: A categorization of existing approaches. In *Datenbanksysteme in Business, Technologie und Web (BTW'07)*, pages 227–241, 2007. (url: http://www.ifi.uzh.ch/dbtg/fileadmin/storage/Glavic/publications/07_BTW_2007_long_version.pdf).

[179] Carole Goble. Position statement: Musings on provenance, workflow and (semantic web) annotations for bioinformatics. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/provenance_workshop_3.doc).

[180] Jennifer Golbeck. Combining provenance with trust in social networks for semantic web content filtering. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 101–108. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_12).

[181] Jennifer Golbeck. Trust on the world wide web: a survey. *Found. Trends Web Sci.*, 1(2):131–197, 2008. (doi: http://dx.doi.org/10.1561/1800000006).

[182] Jennifer Golbeck and James Hendler. A semantic web approach to the provenance challenge. *Concurrency and Computation: Practice and Experience*, 20(5):431–439, 2008. (doi: http://dx.doi.org/10.1002/cpe.1238).

[183] Jennifer Golbeck and Aaron Mannes. Using trust and provenance for content filtering on the semantic web. In *Proceedings of the WWW'06 Workshop on Models of Trust for the Web (MTW'06)*, 2006. (url: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-190/paper02.pdf).

[184] Jose Manuel Gomez-Perez and Oscar Corcho. Problem-solving methods for understanding process executions. *Computing in Science and Engineering*, 10(3):47–52, 2008. (doi: http://doi.ieeecomputersociety.org/10.1109/MCSE.2008.78).

[185] Daniel Goodman. Provenance in dynamically adjusted and partitioned workflows. In *eScience, IEEE International Conference on*, volume 0, pages 39–46, Los Alamitos, CA, USA, 2008. IEEE Computer Society. (doi: http://doi.ieeecomputersociety.org/10.1109/eScience.2008.22).

[186] D. Gotz and M.X. Zhou. Characterizing users' visual analytic activity for insight provenance. In *Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on*, pages 123–130, October 2008. (doi: http://dx.doi.org/10.1109/VAST.2008.4677365).

[187] Todd J Green, Grigoris Karvounarakis, Zachary G Ives, and Val Tannen. Update exchange with mappings and provenance. In *VLDB '07: Proceedings of the 33rd international conference on Very large data bases*, pages 675–686, 2007. (url: http://repository.upenn.edu/cis_reports/763).

[188] Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *PODS '07: Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 31–40, New York, NY, USA, 2007. ACM. (doi: http://doi.acm.org/10.1145/1265530.1265535).

[189] Mark Greenwood, Carole Goble, Robert Stevens, Jun Zhao, Matthew Addis, Darren Marvin, Luc Moreau, and Tom Oinn. Provenance of e-science experiments - experience from bioinformatics. In *Proceedings of the UK OST e-Science second All Hands Meeting 2003 (AHM'03)*, pages 223–226, Nottingham, UK, September 2003. (url: http://www.ecs.soton.ac.uk/~lavm/papers/prov-ahm03.pdf).

[190] Peter C. Griffith, Robert B. Cook, Bruce E. Wilson, Merilyn J. Gentry, Luiz M. Horta, Megan McGroddy, Amy L. Morrell, and Lisa E. Wilcox. Using blackmail, bribery, and guilt to address the tragedy of the virtual intellectual commons. In *Eos Trans. American Geophysical Union, Fall Meeting 2008*, volume 89, 2008. abstract IN11C-1050, (url: http://www.agu.org/cgi-bin/wais?mm=IN11C-1050).

[191] Dennis P. Groth. Information provenance and the knowledge rediscovery problem. In *The Eighth International Conference on Information Visualization*, pages 345–351, Washington, DC, USA, July 2004. IEEE Computer Society. (doi: http://dx.doi.org/10.1109/IV.2004.77).

[192] Dennis P. Groth and Kristy Streefkerk. Provenance and annotation for visual exploration systems. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1500–1510, 2006. (doi: http://doi.ieeecomputersociety.org/10.1109/TVCG.2006.101).

[193] Paul Groth. First opm workshop minutes. http://twiki.ipaw.info/bin/view/challenge/firstopmworkshopminutes, Information Science Institute, USC, July 2008.

[194] Paul Groth, Ewa Deelman, Gideon Juve, Gaurang Mehta, and Bruce Berriman. A pipeline-centric provenance model. In *The 4th Workshop on*

Workflows in Support of Large-Scale Science (WORKS'09), Portland, Oregon, 2009.

[195] Paul Groth, Sheng Jiang, Simon Miles, Steve Munroe, Victor Tan, Sofia Tsasakou, and Luc Moreau. D3.1.1: An architecture for provenance systems. Technical Report http://eprints.ecs.soton.ac.uk/13216/, University of Southampton, November 2006. (url: http://eprints.ecs.soton.ac.uk/13216/).

[196] Paul Groth, Michael Luck, and Luc Moreau. Formalising a protocol for recording provenance in grids. In *Proceedings of the UK OST e-Science second All Hands Meeting 2004 (AHM'04)*, Nottingham, UK, September 2004. (url: http://www.ecs.soton.ac.uk/~lavm/papers/ahm04-groth.pdf).

[197] Paul Groth, Michael Luck, and Luc Moreau. A protocol for recording provenance in service-oriented grids. In *Proceedings of the 8th International Conference on Principles of Distributed Systems (OPODIS'04)*, volume 3544 of *Lecture Notes in Computer Science*, pages 124–139, Grenoble, France, December 2004. Springer-Verlag. (doi: http://dx.doi.org/10.1007/11516798_9).

[198] Paul Groth, Simon Miles, Weijian Fang, Sylvia C. Wong, Klaus-Peter Zauner, and Luc Moreau. Recording and using provenance in a protein compressibility experiment. In *Proceedings of the 14th IEEE International Symposium on High Performance Distributed Computing (HPDC'05)*, pages 201–208, July 2005. (doi: http://dx.doi.org/10.1109/HPDC.2005.1520960).

[199] Paul Groth, Simon Miles, Sanjay Modgil, Nir Oren, Michael Luck, , and Yolanda Gil. Determining the trustworthiness of new electronic contracts. In *Proceedings of the Tenth Annual Workshop on Engineering Societies in the Agents' World, (ESAW'09)*, Utrecht, The Netherlands, November 2009.

[200] Paul Groth, Simon Miles, and Luc Moreau. Preserv: Provenance recording for services. In *Proceedings of the UK OST e-Science second All Hands Meeting 2005 (AHM'05)*, Nottingham,UK, September 2005. (url: http://www.ecs.soton.ac.uk/~lavm/papers/Groth-AHM05.pdf).

[201] Paul Groth, Simon Miles, and Luc Moreau. A model of process documentation to determine provenance in mash-ups. *Transactions on Internet Technology (TOIT)*, 9(1):1–31, 2009. (doi: http://doi.acm.org/10.1145/1462159.1462162).

[202] Paul Groth, Simon Miles, and Steve Munroe. Principles of high quality documentation for provenance: A philosophical discussion. In Luc Moreau and

Ian Foster, editors, *International Provenance and Annotation Workshop (IPAW'06),*, volume 4145 of *Lecture Notes in Computer Science*. Springer, May 2006. (doi: http://dx.doi.org/10.1007/11890850_28).

[203] Paul Groth, Simon Miles, Steve Munroe, Sheng Jiang, Victor Tan, John Ibbotson, and Luc Moreau. D3.2.1: The open provenance specification. Technical report, University of Southampton, November 2006. (url: http://www.gridprovenance.org/deliverables/GRID_PROVENANCE-OpenSpecification-D321-Month24.pdf).

[204] Paul Groth and Luc Moreau. Recording process documentation for provenance. *IEEE Transactions on Parallel and Distributed Systems*, In publication, September 2009. (doi: http://doi.ieeecomputersociety.org/10.1109/TPDS.2008.215).

[205] Paul Groth, Steve Munroe, Simon Miles, and Luc Moreau. *In Lucio Grandinetti (ed.), HPC and Grids in Action (Volume 16 Advances in Parallel Computing)*, chapter Applying the Provenance Data Model to a Bioinformatics Case. IOS Press, January 2008. (url: http://www.ecs.soton.ac.uk/~lavm/papers/hpc08.pdf).

[206] Paul T. Groth. *The Origin of Data: Enabling the Determination of Provenance in Multi-institutional Scientific Systems through the Documentation of Processes.* PhD thesis, Electronics and Computer Science, University of Southampton, 2007. (url: http://eprints.ecs.soton.ac.uk/14649/).

[207] Paul T. Groth. A distributed algorithm for determining the provenance of data. In *Proceedings of the fourth IEEE International Conference on e-Science (e-Science'08)*, 2008. (doi: http://dx.doi.org/10.1109/eScience.2008.38).

[208] Ted Habermann. How can international standards support scientific lineage needs? In *Eos Trans. American Geophysical Union, Fall Meeting 2008*, volume 89, 2008. abstract IN11C-1037, (url: http://www.agu.org/cgi-bin/wais?mm=IN11C-1037).

[209] Christian Halaschek-Wiener, Jennifer Golbeck, Andrew Schain, Michael Grove, Bijan Parsia, and Jim Hendler. Annotation and provenance tracking in semantic web photo libraries. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 82–89. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_10).

[210] Wendy Hall, David De Roure, and Nigel Shadbolt. The evolution of the web and implications for eresearch. *Philosophical Transactions of the Royal So-*

ciety A: Mathematical, Physical and Engineering Sciences, 367(1890):991–1001, 2009. (doi: http://dx.doi.org/10.1098/rsta.2008.0252).

[211] Andreas Harth, Axel Polleres, and Stefan Decker. Towards a social provenance model for the web. In *Workshop on Principles of Provenance (PrOPr)*, Edinburgh, Scotland, 2007. (url: http://sw.deri.org/2007/02/swsepaper/harth-propr.pdf).

[212] Olaf Hartig. Provenance information in the web of data. In *Proceedings of the Linked Data on the Web Workshop (LDOW'09)*, Madrid, Spain, April 2009. (url: http://events.linkeddata.org/ldow2009/papers/ldow2009_paper18.pdf).

[213] Olaf Hartig and Jun Zhao. Using web data provenance for quality assessment. In *Proceedings of the 1st Int. Workshop on the Role of Semantic Web in Provenance Management (SWPM'09) at ISWC*, 2009. (url: http://www.dbis.informatik.hu-berlin.de/fileadmin/research/papers/conferences/2009_swpm_hartig.pdf).

[214] Ragib Hasan, Radu Sion, and Marianne Winslett. Introducing secure provenance: problems and challenges. In *StorageSS '07: Proceedings of the 2007 ACM workshop on Storage security and survivability*, pages 13–18, New York, NY, USA, 2007. ACM. (doi: http://doi.acm.org/10.1145/1314313.1314318).

[215] Ragib Hasan, Radu Sion, and Marianne Winslett. The case of the fake picasso: Preventing history forgery with secure provenance. In Margo I. Seltzer and Richard Wheeler, editors, *Proceedings of 7th USENIX Conference on File and Storage Technologies, FAST 2009*, pages 1–14, San Francisco, Ca, February 2009. (url: http://www.usenix.org/events/fast09/tech/full_papers/hasan/hasan.pdf).

[216] Thomas Heinis and Gustavo Alonso. Efficient lineage tracking for scientific workflows. In *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1007–1018, New York, NY, USA, 2008. ACM. (doi: http://doi.acm.org/10.1145/1376616.1376716).

[217] David A. Holland, Uri Braun, Diana Maclean, Kiran-Kumar Muniswamy-Reddy, and Margo Seltzer. Choosing a data model and query language for provenance. Technical report, Harvard University, 2008. (url: http://www.eecs.harvard.edu/~kiran/pubs/ipaw08.pdf).

[218] David A. Holland, Margo Seltzer, Uri Braun, and Kiran-Kumar Muniswamy-Reddy. Pass-ing the provenance challenge. *Concurrency*

and Computation: Practice and Experience, 20(5), 2008. (doi: http://dx.doi.org/10.1002/cpe.1227).

[219] Bill Howe and Dave Maier. Modeling data product generation. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/position.pdf).

[220] Jiansheng Huang, Ting Chen, AnHai Doan, and Jeffrey F. Naughton. On the provenance of non-answers to queries over extracted data. *Proc. VLDB Endow.*, 1(1):736–747, 2008. (doi: http://doi.acm.org/10.1145/1453856.1453936).

[221] Jingwei Huang and Mark S. Fox. Uncertainty in knowledge provenance. In *The Semantic Web: Research and Applications*, volume 3054 of *Lecture Notes in Computer Science*, pages 372–387. Springer, 2004. (doi: http://dx.doi.org/10.1007/b97867).

[222] Jingwei Huang and Mark S. Fox. Dynamic knowledge provenance. In *Proceedings of Business Agents and Semantic Web Workshop*, 2008. (url: http://www.scientificcommons.org/41069886).

[223] Jinwei Huang and M.S. Fox. Trust judgment in knowledge provenance. In *Database and Expert Systems Applications, 2005. Proceedings. Sixteenth International Workshop on*, pages 524–528, Aug. 2005. (doi: http://dx.doi.org/10.1109/DEXA.2005.193).

[224] Jane Hunter and Kwok Cheung. Provenance explorer-a graphical interface for constructing scientific publication packages from provenance trails. *Int. J. Digit. Libr.*, 7(1):99–107, 2007. (doi: http://dx.doi.org/10.1007/s00799-007-0018-5).

[225] Ian Jackson. Information and informatics in a geological survey - the good, the bad and the ugly. In *Eos Trans. American Geophysical Union, Fall Meeting 2008*, volume 89, 2008. abstract IN11C-1049, (url: http://www.agu.org/cgi-bin/wais?mm=IN11C-1049).

[226] Laura Chiticarius James Cheney and Wang-Chiew Tan. Provenance in databases: Why, how, and where. *Foundations and Trends in Databases*, 1(4):379–474, 2009. (doi: http://dx.doi.org/10.1561/1500000006).

[227] T. J. Jankun-Kelly. Using visualization process graphs to improve visualization exploration. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 78–91. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_10).

[228] Anastasios Kementsietsidis and Min Wang. On the efficiency of provenance queries. In *Data Engineering, International Conference on*, volume 0, pages 1223–1226, Los Alamitos, CA, USA, 2009. IEEE Computer Society. (doi: http://doi.ieeecomputersociety.org/10.1109/ICDE.2009.206).

[229] Imran Khan, Ronald Schroeter, and Jane Hunter. Implementing a secure annotation service. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 212–221. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_22).

[230] Tamas Kifor, Laszlo Varga, Sergio Alvarez, Javier Vazquez-Salceda, and Steven Willmott. Privacy issues of provenance in electronic healthcare record systems. In *First International Workshop on Privacy and Security in Agent-based Collaborative Environments (PSACE2006), AAMAS 2006*, 2006. (url: http://www.gridprovenance.org/publications/EHCR-Prov-Privacy.pdf).

[231] Tamás Kifor, László Z. Varga, Javier Vázquez-Salceda, Sergio Álvarez, Steven Willmott, Simon Miles, and Luc Moreau. Provenance in agent-mediated healthcare systems. *IEEE Intelligent Systems*, 21(6):38–46, Nov/Dec 2006. (doi: http://doi.ieeecomputersociety.org/10.1109/MIS.2006.119).

[232] Jihie Kim, Ewa Deelman, Yolanda Gil, Gaurang Mehta, and Varun Ratnakar. Provenance trails in the wings/pegasus system. *Concurrency and Computation: Practice and Experience*, 20(5):587–597, 2008. (doi: http://dx.doi.org/10.1002/cpe.1228).

[233] Guy K. Kloss and Andreas Schreiber. Provenance implementation in a scientific simulation environment. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 37–45. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_5).

[234] Graham klyne. Contexts for rdf information modelling. (url: http://www.ninebynine.org/RDFNotes/RDFContexts.html).

[235] Christoph Koch. Citations, certificates and object references. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/prov_chicago.pdf).

[236] A. Krenek, J Sitera, J. Chudoba, F. Dvorak, J. Filipovi, J. Kmunicek, L. Matyska, M Mulas, M Ruda, Z. Sustr, S. Campana, E. Molinari, and

D. Rebatto. Experimental evaluation of job provenance in atlas environment. *J. Phys.: Conf Series*, 119, 2007. (doi: http://dx.doi.org/10.1088/1742-6596/119/6/062034).

[237] Ales Krenek, Ludek Matyska, Jirí Sitera, Miroslav Ruda, Frantisek Dvorák, Jiri Filipovic, Zdenek Sustr, and Zdenek Salvet. Job provenance - insight into very large provenance datasets. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 144–151. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_16).

[238] Ales Krenek, Jiri Sitera, Ludek Matyska, Frantisek Dvorak, Milos Mulac, Miroslav Ruda, and Zdenek Salvet. glite job provenance – a job-centric view. *Concurrency and Computation: Practice and Experience*, 20(5):453–462, 2008. (doi: http://dx.doi.org/10.1002/cpe.1252).

[239] Markus Kunde, Henning Bergmeyer, and Andreas Schreiber. Requirements for a provenance visualization component. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 241–252. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_25).

[240] Natalia Kwasnikowska and Jan Van den Bussche. Mapping the nrc dataflow model to the open provenance model. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 3–16. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_3).

[241] David P. Lanter. A lineage meta-database approach toward spatial analytic database optimization. *Cartography and Geographic Information Science*, 20(2):112–121, April 1993. (doi: http://dx.doi.org/10.1559/152304093782610315).

[242] D.P. Lanter. Design of a lineage-based meta-data base for gis. *Cartography and Geographic Information Systems*, 18(4):255–261, 1991.

[243] D.P. Lanter. Lineage in GIS: The problem and a solution. Technical Report 90-6, National Center for Geographic Information and Analysis (NCGIA), UCSB, Santa Barbara, CA, 1991. (url: http://downloads2.esri.com/campus/uploads/library/pdfs/5819.pdf).

[244] D.P. Lanter and R. Essinger. User-centered graphical user interface design for GIS. Technical Report 91-6, National Center for Geographic Information

and Analysis (NCGIA). UCSB, 1991. (url: http://www.ncgia.ucsb.edu/Publications/Tech_Reports/91/91-6.pdf).

[245] Abed Elhamid Lawabni, Changjin Hong, David H. C. Du, and Ahmed H. Tewfik. A novel update propagation module for the data provenance problem: A contemplating vision on realizing data provenance from models to storage. In *MSST '05: Proceedings of the 22nd IEEE / 13th NASA Goddard Conference on Mass Storage Systems and Technologies*, pages 61–69, Washington, DC, USA, 2005. IEEE Computer Society. (doi: http://dx.doi.org/10.1109/MSST.2005.2).

[246] B.N Lawrence, R Lowry, P Miller, H Snaith, and A Woolf. Information in environmental data grids. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1890):1003–1014, 2009. (doi: http://dx.doi.org/10.1098/rsta.2008.0237).

[247] Jonathan Ledlie, Chaki Ng, David A. Holland, Kiran-Kumar Muniswamy-Reddy, Uri Braun, and Margo Seltzer. Provenance-aware sensor data storage. In *Data Engineering Workshops, 2005. 21st International Conference on*, April 2005. (doi: http://dx.doi.org/10.1109/ICDE.2005.270).

[248] T. Lee, S. Bressan, and S. Madnick. Source attribution for querying against semi-structured documents. In *In First Workshop on Web Information and Data Management*, pages 33–39, 1998. (url: http://context2.mit.edu/coin/publications/widm98/widm98.pdf).

[249] Michael Lesk. Data provenance and preservation. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/lesk.txt).

[250] Brian Neil Levine and Marc Liberatore. Dex: Digital evidence provenance supporting reproducibility and comparison. In *Proceedings of the Digital Forensic Research workshop(DFRWS'09)*, 2009. (doi: http://dx.doi.org/10.1016/j.diin.2009.06.011).

[251] Qinglan Li, Alexandros Labrinidis, and Panos K. Chrysanthis. User-centric annotation management for biological data. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 54–61. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_7).

[252] Lauro Lins, David Koop, Erik W. Anderson, Steven P. Callahan, Emanuele Santos, Carlos E. Scheidegger, Juliana Freire, and Cláudio T. Silva. Examining statistics of workflow evolution provenance: A first study. In

SSDBM '08: Proceedings of the 20th international conference on Scientific and Statistical Database Management, pages 573–579, Berlin, Heidelberg, 2008. Springer-Verlag. (doi: http://dx.doi.org/10.1007/978-3-540-69497-7_40).

[253] David T. Liu and Michael J. Franklin. Griddb: a data-centric overlay for scientific grids. In *VLDB '04: Proceedings of the Thirtieth international conference on Very large data bases*, pages 600–611. VLDB Endowment, 2004. (url: http://www.cs.berkeley.edu/~franklin/Papers/griddb_vldb04.pdf).

[254] David B. Lomet. Letter from the editor-in-chief. *IEEE Data Eng. Bull.*, 30(4):1, 2007. (url: http://sites.computer.org/debull/A07dec/dave-let.pdf).

[255] Phillip Lord, Pinar Alper, Chris Wroe, Robert Stevens, Carole Goble, Jun Zhao, Duncan Hull, and Mark Greenwood. The semantic web: Service discovery and provenance in my grid. 2004. (url: http://lists.w3.org/Archives/Public/public-swls-ws/2004Sep/att-0016/semantic_web_for_life_sciences_position.pdf).

[256] Bertram Ludaescher, Norbert Podhorszki, Ilkay Altintas, Shawn Bowers, and Timothy M. McPhillips. From computation models to models of provenance: the rws approach. *Concurrency and Computation: Practice and Experience*, 20(5):519–529, 2008. (doi: http://dx.doi.org/10.1002/cpe.1234).

[257] Clifford A. Lynch. When documents deceive: trust and provenance as new factors for information retrieval in a tangled web. *Journal of the American Society for Information Science and Technology*, 52(1):12–17, 2001. (doi: /10.1002/1532-2890(2000)52:1<12::AID-ASI1062>3.3.CO;2-M).

[258] Chris Lynnes, Gregory Leptoukh, Stephen W. Berrick, Suhung Shen, Ana I. Prados, Peter Fox, Wenli Yang, M Min, Daniel Holloway, and Yonsook Enloe. Provenance in data interoperability for multi-sensor intercomparison. In *Eos Trans. American Geophysical Union, Fall Meeting 2008*, volume 89, 2008. abstract IN11C-1041, (url: http://www.agu.org/cgi-bin/wais?mm=IN11C-1041).

[259] Allan MacKenzie-Graham, Arash Payan, Ivo D. Dinov, John D. Van Horn, and Arthur W. Toga. Neuroimaging data provenance using the loni pipeline workflow environment. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*,

pages 208–220. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_22).

[260] Stewart P. Macleod, Casey L. Kiernan, and WA) Rajarajan, Vij (Issaquah. Data lineage data type. United States Patent 6434558, United States Patent, 2002. (url: http://www.freepatentsonline.com/6434558.html).

[261] Bob Mann. Some data derivation and provenance issues in astronomy. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/mann.ps).

[262] Arunprasad P. Marathe. Tracing lineage of array data. In *Proceedings of the 13th International Conference on Scientific and Statistical Database Management (SSDBM'01)*, pages 69–78, Fairfax, VA, July 2001. (doi: http://doi.ieeecomputersociety.org/10.1109/SSDM.2001.938539).

[263] Arunprasad P. Marathe. Tracing lineage of array data. *Journal of Intelligent Information Systems*, 17(2-3):193–214, 2001. (doi: http://dx.doi.org/10.1023/A:1012857830230).

[264] Daniel W. Margo and Margo Seltzer. The case for browser provenance. In James Cheney, editor, *TAPP'09: First workshop on on Theory and practice of provenance*, San Francisco, CA, February 2009. USENIX Association. (url: http://www.usenix.org/event/tapp09/tech/full_papers/margo/margo.pdf).

[265] A. Marins, M. A. Casanova, and K. Breitman A. Furtado. Modeling provenance for semantic desktop applications. In *Anais do XXVII Congresso da SBC (SBC'07)*, pages 2100–2112, Rio de Janeiro, Brazil, jul 2007. (url: http://cidoc.ics.forth.gr/docs/ModelingProvenanceforSemanticDesktopApplications.pdf).

[266] Chris Martin, Mohammed H. Haji, Peter M. Dew, Mike Pilling, and Peter K. Jimack. Semantically-enhanced model-experiment-evaluation processes (semeeps) within the atmospheric chemistry community. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 293–308. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_29).

[267] Chris J Martin, Mohammed H Haji, Peter M Dew, Michael J Pilling, and Peter K Jimack. Semantically enhanced provenance capture for chamber model development with a master chemical mechanism. *Philosophical*

Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367(1890):987–990, 2009. (doi: http://dx.doi.org/10.1098/rsta.2008.0168).

[268] Matyska. Job tracking on a grid - the logging and bookkeeping and job provenance services. Technical report, CESNET, 2007. (url: http://www.cesnet.cz/doc/techzpravy/2007/grid-job-tracking/).

[269] Michael McCann and Kevin Gomes. Oceanographic data provenance tracking with the shore side data system. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 309–322. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_30).

[270] Deborah L. McGuinness and Paulo Pinheiro da Silva. Infrastructure for web explanations. In *International Semantic Web Conference*, pages 113–129, 2003. (doi: http://dx.doi.org/10.1007/b14287).

[271] Deborah L. McGuinness and Paulo Pinheiro da Silva. Explaining answers from the semantic web: the inference web approach. *J. Web Sem.*, 1(4):397–413, 2004. (doi: http://dx.doi.org/10.1016/j.websem.2004.06.002).

[272] Deborah L. McGuinness, Peter Fox, Paulo Pinheiro da Silva, Stephan Zednik, Nicholas Del Rio, Li Ding, Patrick West, and Cynthia Chang. Annotating and embedding provenance in science data repositories to enable next generation science applications. In *Eos Trans. American Geophysical Union, Fall Meeting 2008*, volume 89, 2008. abstract IN11C-1052, (url: http://www.agu.org/cgi-bin/wais?mm=IN11C-1052).

[273] Timothy Mcphillips, Shawn Bowers, and Bertram Ludaescher. Collection-oriented scientific workflows for integrating and analyzing biological data. In *Data Integration in the Life Sciences*, pages 248–263, 2006. (doi: http://dx.doi.org/10.1007/11799511_23).

[274] Anton Michlmayr, Florian Rosenberg, Philipp Leitner, and Schahram Dustdar. Service provenance in qos-aware web service runtimes. In *Proceedings of the 7th IEEE International Conference on Web Services (ISWC'09)*, Los Angeles, Ca, July 2009. (url: http://www.infosys.tuwien.ac.at/Staff/rosenberg/papers/icws2009.pdf).

[275] Simon Miles. Agent-oriented data curation in bioinformatics. In *Proceedings of Workshop on Multi-Agent Systems in Medicine, Computational Biology, and Bioinformatics (MAS*BioMed'05)*, July 2005. (url: http://eprints.ecs.soton.ac.uk/10853/).

[276] Simon Miles. Electronically querying for the provenance of entities. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 184–192. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_19).

[277] Simon Miles. Technical summary of the second provenance challenge workshop. http://twiki.ipaw.info/bin/view/challenge/secondworkshopminutes, King's College, July 2007.

[278] Simon Miles, Ewa Deelman, Paul Groth, Karan Vahi, Gaurang Mehta, and Luc Moreau. Connecting scientific data to scientific experiments with provenance. In *Proceedings of the third IEEE International Conference on e-Science and Grid Computing (e-Science'07)*, pages 179–186, Bangalore, India, December 2007. (doi: http://doi.ieeecomputersociety.org/10.1109/E-SCIENCE.2007.22).

[279] Simon Miles, Paul Groth, Miguel Branco, and Luc Moreau. The requirements of using provenance in e-science experiments. *Journal of Grid Computing*, 5(1):1–25, 2007. (doi: http://dx.doi.org/10.1007/s10723-006-9055-3).

[280] Simon Miles, Paul Groth, Ewa Deelman, Karan Vahi, Gaurang Mehta, and Luc Moreau. Provenance: The bridge between experiments and data. *Computing in Science and Engineering*, 10(3):38–46, May/June 2008. (doi: http://doi.ieeecomputersociety.org/10.1109/MCSE.2008.82).

[281] Simon Miles, Paul Groth, and Michael Luck. Handling mitigating circumstances for electronic contracts. In *Proceedings of the AISB 2008 Symposium on Behaviour Regulation in Multi-agent Systems*, pages 37–42, Aberdeen, UK, April 2008. The Society for the Study of Artificial Intelligence and Simulation of Behaviour. (url: http://calcium.dcs.kcl.ac.uk/1283/).

[282] Simon Miles, Paul Groth, Steve Munroe, Sheng Jiang, Thibaut Assandri, and Luc Moreau. Extracting causal graphs from an open provenance data model. *Concurrency and Computation: Practice and Experience*, 20(5):577–586, 2008. (doi: http://dx.doi.org/10.1002/cpe.1236).

[283] Simon Miles, Paul Groth, Steve Munroe, Michael Luck, and Luc Moreau. Agentprime: Adapting mas designs to build confidence. In *Agent-Oriented Software Engineering (AOSE'07)*, volume 4951 of *Lecture Notes in Computer Science*. Springer, 2007. (doi: http://dx.doi.org/10.1007/978-3-540-79488-2_3).

[284] Simon Miles, Paul Groth, Steve Munroe, and Luc Moreau. Prime: A methodology for developing provenance-aware applications. *ACM Transactions on Software Engineering and Methodology*, 2009.

[285] Simon Miles, Steve Munroe, Michael Luck, and Luc Moreau. Modelling the provenance of data in autonomous systems. In *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'07)*, pages 1–8, New York, NY, USA, 2007. ACM. (doi: http://doi.acm.org/10.1145/1329125.1329185).

[286] Simon Miles, Sylvia C. Wong, Weijian Fang, Paul Groth, Klaus-Peter Zauner, and Luc Moreau. Provenance-based validation of e-science experiments. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(1):28–38, 2007. (doi: http://dx.doi.org/10.1016/j.websem.2006.11.003).

[287] Archan Misra, Marion Blount, Anastasios Kementsietsidis, Daby M. Sow, and Min Wang. Advances and challenges for scalable provenance in stream processing systems. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 253–265. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_26).

[288] Paolo Missier, Khalid Belhajjame, Jun Zhao, Marco Roos, and Carole A. Goble. Data lineage model for taverna workflows with lightweight annotation requirements. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 17–30. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_4).

[289] Paolo Missier, Suzanne M. Embury, and Richard Stapenhurst. Exploiting provenance to make sense of automated decisions in scientific workflows. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 174–185. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_19).

[290] Luc Moreau. Usage of 'provenance': A tower of babel. towards a concept map — position paper for the microsoft life cycle seminar, mountain view, july 10, 2006. Technical report, University of Southampton, June 2006. (url: http://www.ecs.soton.ac.uk/~lavm/papers/babel.pdf).

[291] Luc Moreau and Ian Foster, editors. *Provenance and Annotation of Data — International Provenance and Annotation Workshop, IPAW 2006*, volume

4145 of *Lecture Notes in Computer Science*. Springer, May 2006. (doi: http://dx.doi.org/10.1007/11890850).

[292] Luc Moreau, Juliana Freire, Joe Futrelle, Robert E. McGrath, Jim Myers, and Patrick Paulson. The open provenance model (v1.00). Technical report, University of Southampton, December 2007. (url: http://eprints.ecs.soton.ac.uk/14979/).

[293] Luc Moreau, Juliana Freire, Joe Futrelle, Robert E. McGrath, Jim Myers, and Patrick Paulson. The open provenance model: An overview. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 323–326. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_31).

[294] Luc Moreau, Paul Groth, Simon Miles, Javier Vazquez, John Ibbotson, Sheng Jiang, Steve Munroe, Omer Rana, Andreas Schreiber, Victor Tan, and Laszlo Varga. The provenance of electronic data. *Communications of the ACM*, 51(4):52–58, April 2008. (doi: http://doi.acm.org/10.1145/1330311.1330323).

[295] Luc Moreau and John Ibbotson. Standardisation of provenance systems in service oriented architectures — white paper. Technical report, University of Southampton, 2006. (url: http://eprints.ecs.soton.ac.uk/12198/).

[296] Luc Moreau, Natalia Kwasnikowska, and Jan Van den Bussche. The foundations of the open provenance model. Technical report, University of Southampton, April 2009. (url: http://eprints.ecs.soton.ac.uk/17282/).

[297] Luc Moreau and Bertram Ludaescher, editors. *Special Issue on the First Provenance Challenge*, volume 20. Wiley, April 2008. (doi: http://dx.doi.org/10.1002/cpe.1233).

[298] Luc Moreau, Bertram Ludaescher, Ilkay Altintas, Roger S. Barga, Shawn Bowers, Steven Callahan, George Chin Jr., Ben Clifford, Shirley Cohen, Sarah Cohen-Boulakia, Susan Davidson, Ewa Deelman, Luciano Digiampietri, Ian Foster, Juliana Freire, James Frew, Joe Futrelle, Tara Gibson, Yolanda Gil, Carole Goble, Jennifer Golbeck, Paul Groth, David A. Holland, Sheng Jiang, Jihie Kim, David Koop, Ales Krenek, Timothy McPhillips, Gaurang Mehta, Simon Miles, Dominic Metzger, Steve Munroe, Jim Myers, Beth Plale, Norbert Podhorszki, Varun Ratnakar, Emanuele Santos, Carlos Scheidegger, Karen Schuchardt, Margo Seltzer, Yogesh L. Simmhan, Claudio Silva, Peter Slaughter, Eric Stephan, Robert Stevens, Daniele Turi, Huy Vo, Mike Wilde, Jun Zhao, and Yong Zhao. The first

provenance challenge. *Concurrency and Computation: Practice and Experience*, 20(5):409–418, April 2008. (doi: http://dx.doi.org/10.1002/cpe.1233).

[299] Luc Moreau (Editor), Beth Plale, Simon Miles, Carole Goble, Paolo Missier, Roger Barga, Yogesh Simmhan, Joe Futrelle, Robert McGrath, Jim Myers, Patrick Paulson, Shawn Bowers, Bertram Ludaescher, Natalia Kwasnikowska, Jan Van den Bussche, Tommy Ellkvist, Juliana Freire, and Paul Groth. The open provenance model (v1.01). Technical report, University of Southampton, July 2008. (url: http://eprints.ecs.soton.ac.uk/16148/1/opm-v1.01.pdf).

[300] Pierre Mouallem, Roselyne Barreto, Scott Klasky, Norbert Podhorszki, and Mladen Vouk. Tracking files in the kepler provenance framework. In *Proceedings of 21st International Conference on Scientific and Statistical Database Management (SSDBM'09)*, pages 273–282, New Orleans, LA, USA, 2009. (doi: http://dx.doi.org/10.1007/978-3-642-02279-1_21).

[301] K. Muniswamy-Reddy, D. Holland, Uri. Braun, and Margo. Seltzer. Provenance-aware storage systems. In *ATEC '06: Proceedings of the annual conference on USENIX '06 Annual Technical Conference*, pages 43–56, Berkeley, CA, USA, June 2006. USENIX Association. (url: http://www.usenix.org/events/usenix06/tech/muniswamy-reddy.html).

[302] Kiran-Kumar Muniswamy-Reddy, Uri Braun, David A. Holland, Peter Macko, Diana Maclean, Daniel Margo, Margo Seltzer, and Robin Smogor. Layering in provenance-aware storage systems. In *Proceedings of 2009 USENIX Annual Technical Conference*, San Diego, CA, June 2009.

[303] Kiran-Kumar Muniswamy-Reddy, Peter Macko, and Margo Seltzer. Making a cloud provenance-aware. In James Cheney, editor, *TAPP'09: First workshop on on Theory and practice of provenance*, San Francisco, CA, February 2009. USENIX Association. (url: http://www.usenix.org/event/tapp09/tech/full_papers/muniswamy-reddy/muniswamy-reddy.pdf).

[304] Steve Munroe, Simon Miles, Luc Moreau, and Javier Vázquez-Salceda. PrIMe: A software engineering methodology for developing provenance-aware applications. In *ACM Digital Proceedings of the Software Engineering and Middleware Workshop (SEM'06)*, pages 39–46, New York, NY, USA, 2006. ACM. (doi: http://doi.acm.org/10.1145/1210525.1210535).

[305] Michi Mutsuzaki, Martin Theobald, Ander de Keijzer, Jennifer Widom, Parag Agrawal, Omar Benjelloun, Anish Das Sarma, Raghotham Murthy, and Tomoe Sugihara. Trio-one: Layering uncertainty and lineage on a

conventional dbms. In *Proc. of CIDR conference (system demonstration)*, 2007. (url: http://ilpubs.stanford.edu:8090/805/).

[306] J. Myers, C. Pancerella, C. Lansing, K. Schuchardt, and B. Didier. Multi-scale science, supporting emerging practice with semantically derived provenance. In *Proceedings of the ISWC 2003 Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data*, Sanibel Island, Florida, October 2003. (url: http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-83/prov_1.pdf).

[307] James Myers. Design constraints for scientific annotation systems. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/SAM1.Provenance.position.doc).

[308] James D. Myers, Joe Futrelle, Jeff Gaynor, Joel Plutchak, Peter Bajcsy, Jason Kastner, Kailash Kotwani, Jong Sung Lee, Luigi Marini, Rob Kooper, Robert E. McGrath, Terry McLaren, Alejandro Rodrguez, and Yong Liu. Embedding data within knowledge spaces. In *UK e-sicence 2008*, 2008. (url: http://arxiv.org/abs/0902.0744).

[309] J.D. Myers, A.R. Chappell, M. Elder, A. Geist, and J. Schwidder. Re-integrating the research record. *IEEE Computing in Science and Engineering*, 5(3):44–50, 2003. (doi: http://dx.doi.org/10.1109/MCISE.2003.1196306).

[310] Meiyappan Nagappan and Mladen A. Vouk. A model for sharing of confidential provenance information in a query based system. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 62–69. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_8).

[311] Helen Nissenbaum. Computing and accountability. *Commun. ACM*, 37(1):72–80, 1994. (doi: http://doi.acm.org/10.1145/175222.175228).

[312] John Ockerbloom. Copyright and provenance: Some practical problems. *IEEE Data Eng. Bull.*, 30(4):51–58, 2007. (url: http://sites.computer.org/debull/A07dec/ockerbloom.pdf).

[313] The open provenance web site. http://openprovenance.org, August 2008.

[314] Leon J. Osterweil, Lori A. Clarke, Aaron M. Ellison, Rodion Podorozhny, Alexander Wise, Emery Boose, and Julian Hadley. Experience in using a process language to define scientific workflow and generate dataset provenance. In *SIGSOFT '08/FSE-16: Proceedings of the 16th ACM*

SIGSOFT International Symposium on Foundations of software engineering, pages 319–329, New York, NY, USA, 2008. ACM. (doi: http://doi.acm.org/10.1145/1453101.1453147).

[315] Carmen Pancerella, Jim Myers, and Larry Rahn. Data provenance in the cmcs. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/ProvenanceWorkshopCMCS.pdf).

[316] Unkyu Park and John Heidemann. Provenance in sensornet republishing. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 280–292. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_28).

[317] Pasoa: Provenance aware service oriented architecture. www.pasoa.org, 2004.

[318] Dave Pearson. The grid: Requirements for establishing the provenance of derived data. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/Provenance_Requirements.doc).

[319] P. Pediaditis, G. Flouris, I. Fundulak, and V. Christophides. On explicit provenance management in rdf/s graphs. In James Cheney, editor, *TAPP'09: First workshop on on Theory and practice of provenance*, San Francisco, CA, February 2009. USENIX Association. (url: http://www.usenix.org/event/tapp09/tech/full_papers/pediaditis/pediaditis.pdf).

[320] Lorna Philip, Alison Chorley, John Farrington, and Pete Edwards. Data provenance, evidence-based policy assessment, and e-social science. In *Third International Conference on e-Social Science*, October 2007. (url: http://www.scientificcommons.org/40739576).

[321] Nicolas Prat and Stuart Madnick. Measuring data believability: A provenance approach. In *Proceedings of the 41st Hawaii International Conference on System Sciences - 2008*. IEEE Computer Society, 2008. (doi: http://doi.ieeecomputersociety.org/10.1109/HICSS.2008.243).

[322] Enabling and supporting provenance in grids for complex problems. www.gridprovenance.org, 2005.

[323] Shrija Rajbhandari, Arnaud Contes, Omer F.Rana, Vikas Deora, and Ian Wootten. Establishing workflow trust using provenance information. In *1st IEEE International Workshop on Modelling Autonomic Communications Environments (MACE 2006)*, October 2006. (url: http://www.gridprovenance.org/publications/manweek-ranaetal.pdf).

[324] Shrija Rajbhandari, Arnaud Contes, Omer F. Rana, Vikas Deora, and Ian Wootten. Trust assessment using provenance in service oriented applications. In *EDOCW '06: Proceedings of the 10th IEEE on International Enterprise Distributed Object Computing Conference Workshops*, page 65, Washington, DC, USA, 2006. IEEE Computer Society. (doi: http://dx.doi.org/10.1109/EDOCW.2006.70).

[325] Shrija Rajbhandari, Omer F. Rana, and Ian Wootten. A fuzzy model for calculating workflow trust using provenance data. In *MG '08: Proceedings of the 15th ACM Mardi Gras conference*, pages 1–8, New York, NY, USA, 2008. ACM. (doi: http://doi.acm.org/10.1145/1341811.1341823).

[326] Shrija Rajbhandari and David Walker. Support for provenance in a service-based computing grid. In *Proceedings of the UK OST e-Science second All Hands Meeting 2004 (AHM'04)*, Nottingham, UK, September 2004. (url: http://www.wesc.ac.uk/resources/publications/pdf/AHM04/194.pdf).

[327] Shrija Rajbhandari and David W. Walker. Incorporating provenance in service oriented architecture. In *NWESP '06: Proceedings of the International Conference on Next Generation Web Services Practices*, pages 33–40, Washington, DC, USA, 2006. IEEE Computer Society. (doi: http://dx.doi.org/10.1109/NWESP.2006.18).

[328] Shrija Rajbhandari, Ian Wootten, Ali Shaikh Ali, and Omer F. Rana. Evaluating provenance-based trust for scientific workflows. In *CCGRID '06: Proceedings of the Sixth IEEE International Symposium on Cluster Computing and the Grid*, pages 365–372, Washington, DC, USA, 2006. IEEE Computer Society. (doi: http://dx.doi.org/10.1109/CCGRID.2006.43).

[329] Sudha Ram and Jun Liu. Understanding the semantics of data provenance to support active conceptual modeling. In *Active Conceptual Modeling of Learning*, Lecture Notes in Computer Science, pages 17–29, 2008. (doi: http://dx.doi.org/10.1007/978-3-540-77503-4_3).

[330] Sudha Ram, Jun Liu, and Regi Thomas George. PROMS: A system for harvesting and managing data provenance. In *WITS 2006*, December 2006. (url: http://kartik.eller.arizona.edu/WITS_DEMO_final.pdf).

[331] Sudha Ram, Jun Liu, Nirav Merchant, Terrill Yuhas, and Patty Jansma. Toward developing a provenance ontology for biological images. In *Eighth Annual Bio-Ontologies Workshop*, 2005. (url: http://kartik.eller.arizona.edu/Abstract4_29.doc).

[332] Christopher Ré and Dan Suciu. Approximate lineage for probabilistic databases. *Proc. VLDB Endow.*, 1(1):797–808, 2008. (doi: http://doi.acm.org/10.1145/1453856.1453943).

[333] Alberto Reggiori, Dirk-Willem van Gulik, and Zavisa Bjelogrlic. Indexing and retrieving semantic web resources: the rdfstore model. In *SWAD-Europe Workshop on Semantic Web Storage and Retrieval*, Amsterdam, Netherlands, November 2003. (url: http://www.w3.org/2001/sw/Europe/events/20031113-storage/positions/asemantics.html).

[334] Christine F. Reilly and Jeffrey F. Naughton. Exploring provenance in a distributed job execution system. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 237–245. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_24).

[335] Christine F. Reilly and Jeffrey F. Naughton. Transparently gathering provenance with provenance aware condor. In James Cheney, editor, *TAPP'09: First workshop on on Theory and practice of provenance*, San Francisco, CA, February 2009. USENIX Association. (url: http://www.usenix.org/event/tapp09/tech/full_papers/reilly/reilly.pdf).

[336] Nicholas Del Rio and Paulo Pinheiro da Silva. Probe-it! visualization support for provenance. In *Advances in Visual Computing, Third International Symposium, ISVC 2007, Lake Tahoe, NV, USA, November 26-28, 2007, Proceedings, Part II*, volume 4842 of *Lecture Notes in Computer Science*, pages 732–741. Springer, 2007. (doi: http://dx.doi.org/10.1007/978-3-540-76856-2_72).

[337] Nicholas Del Rio, Paulo Pinheiro da Silva, Ann Q. Gates, and Leonardo Salayandia. Semantic annotation of maps through knowledge provenance. In *GeoSpatial Semantics, Second International Conference, GeoS 2007, Mexico City, Mexico, November 29-30, 2007, Proceedings*, volume 4853 of *Lecture Notes in Computer Science*, pages 20–35, 2007. (doi: http://dx.doi.org/10.1007/978-3-540-76876-0_2).

[338] Arnon Rosenthal, Len Seligman, Adriane Chapman, and Barbara Blaustein. Scalable access controls for lineage. In James Cheney, editor, *TAPP'09: First workshop on on Theory and practice of provenance*, San Francisco, CA, February 2009. USENIX Association. (url: http://www.usenix.org/event/tapp09/tech/full_papers/rosenthal/rosenthal.pdf).

[339] Seth Russell. Quads. (url: http://robustai.net/sailor/grammar/Quads.html).

[340] Paul Ruth, Dongyan Xu, Bharat K. Bhargava, and Fred Regnier. E-notebook middleware for accountability and reputation based trust in distributed data sharing communities. In *Proceedings 2nd International Conference on Trust Management (iTrust'04)*, volume 2995 of *Lecture Notes in Computer Science*, pages 161–175. Springer, 2004. (doi: http://dx.doi.org/10.1007/b96545).

[341] H. Sabaa and B. Panda. Data authentication and provenance management. In *Digital Information Management, 2007. ICDIM '07. 2nd International Conference on*, volume 1, pages 309–314, Lyon, France, October 2007. (doi: http://dx.doi.org/10.1109/ICDIM.2007.4444241).

[342] Satya S. Sahoo, Roger S. Barga, Jonathan Goldstein, and Amit P. Sheth. Provenance algebra and materialized view-based provenance management. Technical Report 76523/tr-2008-170, Microsoft Research, 2008. (url: http://research.microsoft.com/pubs/76523/tr-2008-170.pdf).

[343] Satya S. Sahoo, Amit Sheth, and Cory Henson. Semantic provenance for escience: Managing the deluge of scientific data. *Internet Computing, IEEE*, 12(4):46–54, July-Aug 2008. (doi: http://dx.doi.org/10.1109/MIC.2008.86).

[344] Joel Saltz. Data Provenance. At [144], October 2002. (url: http://people.cs.uchicago.edu/~yongzh/papers/ProvenanceJS10-02.doc).

[345] Emanuele Santos, Lauro Lins, James P. Ahrens, Juliana Freire, and Cláudio T. Silva. A first study on clustering collections of workflow graphs. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 160–173. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_18).

[346] C. Sar and P. Cao. Lineage file system. Technical report, Stanford University, 2005. (url: http://theory.stanford.edu/~cao/lineage).

[347] Anish Das Sarma, Martin Theobald, and Jennifer Widom. Exploiting lineage for confidence computation in uncertain and probabilistic databases. Technical Report 2007-15, Stanford InfoLab, March 2007. (url: http://ilpubs.stanford.edu:8090/800/).

[348] Carlos Scheidegger, David Koop, Emanuele Santos, Huy Vo, Steven Callahan, Juliana Freire, and Claudio Silva. Tackling the provenance challenge one layer at a time. *Concurrency and Computation: Practice and Experience*, 20(5):473–483, 2008. (doi: http://dx.doi.org/10.1002/cpe.1237).

[349] Carlos Scheidegger, David Koop, Huy Vo, Juliana Freire, and Claudio Silva. Querying and creating visualizations by analogy. *IEEE Transactions on Visualization and Computer Graphics*, 2007. (doi: http://doi.ieeecomputersociety.org/10.1109/TVCG.2007.70584).

[350] Carlos E. Scheidegger, Huy T. Vo, David Koop, Juliana Freire, and Claudio T. Silva. Querying and re-using workflows with vistrails. In *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1251–1254, New York, NY, USA, 2008. ACM. (doi: http://doi.acm.org/10.1145/1376616.1376747).

[351] Karen Schuchardt, Tara Gibson, Eric Stephan, and George Chin, Jr. Applying content management to automated provenance capture. *Concurrency and Computation: Practice and Experience*, 20:541–554, 2008. (doi: http://dx.doi.org/10.1002/cpe.1230).

[352] Second challenge team contributions. http://twiki.ipaw.info/bin/view/Challenge/ParticipatingTea June 2007.

[353] Sam Shah, Craig A. N. Soules, Gregory R. Ganger, and Brian D. Noble. Using provenance to aid in personal file search. In *ATC'07: 2007 USENIX Annual Technical Conference on Proceedings of the USENIX Annual Technical Conference*, pages 1–14, Berkeley, CA, USA, 2007. USENIX Association. (url: http://www.pdl.cmu.edu/PDL-FTP/ABN/usenix07.pdf).

[354] Eddie C. Shek and Richard R. Muntz. Exploiting data lineage for parallel optimization in extensible dbmss. In *ICDE '99: Proceedings of the 15th International Conference on Data Engineering*, page 256, Washington, DC, USA, 1999. IEEE Computer Society. (doi: http://doi.ieeecomputersociety.org/10.1109/ICDE.1999.754936).

[355] Chris A. Silles and Andrew R. Runnalls. Provenance tracking in cxxr. In *The R User Conference 2009*, Agrocampus-Ouest, Rennes, France, July 2009. (url: http://www.agrocampus-ouest.fr/math/useR-2009/abstracts/pdf/Silles_Runnalls.pdf).

[356] Claudio Silva, Juliana Freire, and Steven P. Callahan. Provenance for visualizations: Reproducibility and beyond. *Computing in Science and Engineering*, 9(5):82–89, 2007. (doi: http://doi.ieeecomputersociety.org/10.1109/MCSE.2007.106).

[357] Claudio T. Silva and Joel E. Tohline. Computational provenance. *Computing in Science and Engineering*, 10(3):9–10, 2008. (doi: http://doi.ieeecomputersociety.org/10.1109/MCSE.2008.71).

[358] Yogesh Simmhan. *Provenance Framework in Support of Data Quality Estimation*. PhD thesis, University of Indiana, 2007. (url: http://gradworks.umi.com/32/97/3297094.html).

[359] Yogesh L. Simmhan, Roger S. Barga, and Catharine van Ingen. Automatic provenance recording for scientific data using trident. In *Eos Trans. American Geophysical Union, Fall Meeting 2008*, volume 89, 2008. abstract IN11C-1048, (url: http://www.agu.org/cgi-bin/wais?mm=IN11C-1048).

[360] Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. A survey of data provenance in e-science. *SIGMOD Record*, 34(3):31–36, 2005. (doi: http://doi.acm.org/10.1145/1084805.1084812).

[361] Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. A framework for collecting provenance in data-centric scientific workflows. In *International Conference on Web Service (ICWS'06)*, pages 427–436, Washington, DC, USA, 2006. IEEE Computer Society. (doi: http://dx.doi.org/10.1109/ICWS.2006.5).

[362] Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. Karma2: Provenance management for data driven workflows. *International Journal of Web Services Research*, 5(2), 2008. (url: http://www.cs.indiana.edu/~plale/papers/Simmhan-JWSR-07.pdf).

[363] Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. Querying capabilities of the karma provenance framework. *Concurrency and Computation: Practice and Experience*, 20(5):441–451, 2008. (doi: http://dx.doi.org/10.1002/cpe.1229).

[364] Yogesh L. Simmhan, Beth Plale, Dennis Gannon, and Suresh Marru. Performance evaluation of the karma provenance framework for scientific workflows. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 222–236. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_23).

[365] Roger W. Smith. Sharing data resources benefits owners as well as miners. In *Eos Trans. American Geophysical Union, Fall Meeting 2008*, volume 89, 2008. abstract IN11C-1051, (url: http://www.agu.org/cgi-bin/wais?mm=IN11C-1051).

[366] Issam Souilah, Adrian Francalanza, and Vladimiro Sassone. A formal model of provenance in distributed systems. In James Cheney, editor, *TAPP'09: First workshop on on Theory and practice of provenance*, San Francisco, CA, February 2009. USENIX Association. (url: http://www.usenix.org/event/tapp09/tech/full_papers/souilah/souilah.pdf).

[367] Laurent Spery, Christophe Claramunt, and Thérèse Libourel. A lineage metadata model for the temporal management of a cadastre application. In *DEXA '99: Proceedings of the 10th International Workshop on Database and Expert Systems Applications*, page 466, Washington, DC, USA, 1999. IEEE Computer Society. (url: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00795211).

[368] Laurent Spery, Christophe Claramunt, and Thérèse Libourel. A spatio-temporal model for the manipulation of lineage metadata. *Geoinformatica*, 5(1):51–70, 2001. (doi: http://dx.doi.org/10.1023/A:1011459921552).

[369] R. Spillane, R. Sears, C. Yalamanchill, S. Gaikwad, M. Chinni, and E. Zadok. Story book: an efficient extensible provenance framework. In James Cheney, editor, *TAPP'09: First workshop on on Theory and practice of provenance*, San Francisco, CA, February 2009. USENIX Association. (url: http://www.usenix.org/event/tapp09/tech/full_papers/spillane/spillane.pdf).

[370] Divesh Srivastava and Yannis Velegrakis. Intensional associations between data and metadata. In *SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 401–412, New York, NY, USA, 2007. ACM. (doi: http://doi.acm.org/10.1145/1247480.1247526).

[371] Robert Stevens, Jun Zhao, and Carole Goble. Using provenance to manage knowledge of in silico experiments. *Briefing in Bioinformatics*, 8(3):183–194, 2007. (doi: http://dx.doi.org/10.1093/bib/bbm015).

[372] Igor Suarez-Sola, Alisdair Davey, and Joseph A. Hourcle. What are we tracking ... and why? In *Eos Trans. AGU*, volume 89, 2008. abstract IN11C-1047, (url: http://www.agu.org/cgi-bin/wais?mm=IN11C-1047).

[373] Peng Sun, Ziyang Liu, Susan B. Davidson, and Yi Chen. Detecting and resolving unsound workflow views for correct provenance analysis. In *SIGMOD '09: Proceedings of the 35th SIGMOD international conference on Management of data*, pages 549–562, New York, NY, USA, 2009. ACM. (doi: http://doi.acm.org/10.1145/1559845.1559903).

[374] Amril Syalim, Yoshiaki Hori, and Kouichi Sakurai. Grouping provenance information to improve efficiency of access control. In *Third International Conference and Workshops on Advances in Information Security and Assurance (ISA'09)*, volume 5576 of *Lecture Notes in Computer Science*, pages 51–59, 2009. (doi: http://dx.doi.org/10.1007/978-3-642-02617-1_6).

[375] Martin Szomszor and Luc Moreau. Recording and reasoning over data provenance in web and grid services. In *International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE'03)*, volume 2888 of *Lecture Notes in Computer Science*, pages 603–620, Catania, Sicily, Italy, November 2003. (doi: http://dx.doi.org/10.1007/b94348).

[376] Tara D. Talbott, Karen L. Schuchardt, Eric G. Stephan, and James D. Myers. Mapping physical formats to logical models to extract data and metadata: The defuddle parsing engine. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 73–81. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_9).

[377] Victor Tan, Paul Groth, Simon Miles, Sheng Jiang, Steve Munroe, Sofia Tsasakou, and Luc Moreau. Security issues in a soa-based provenance system. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop (IPAW'06)*, volume 4145 of *Lecture Notes in Computer Science*, pages 203–211, Chicago, Illinois, May 2006. Springer-Verlag. (doi: http://dx.doi.org/10.1007/11890850_21).

[378] Wang-Chiew Tan. *Data Annotations, Provenance, and Archiving*. PhD thesis, U Penn., Philadelphia, PA, USA, 2002. Supervisor-Buneman, Peter and Supervisor-Khanna, Sanjeev, (url: http://proquest.umi.com/pqdlink?did=765108921&Fmt=7&clientId=79356&RQT=309&VName=PQD).

[379] Wang Chiew Tan. Research problems in data provenance. *IEEE Data Eng. Bull.*, 27(4):45–52, 2004. (url: http://db.cs.ucsc.edu/node/216).

[380] Wang-Chiew Tan. Provenance in databases: Past, current, and future. *Bulletin of the Technical Committee on Data Engineering*, 30(4):3–12, December 2007. (url: ftp://ftp.research.microsoft.com/pub/debull/A07dec/wang-chiew.pdf).

[381] Val Tannen. Provenance for database transformations. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, page 1. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_1).

[382] The Southampton Provenance Team. Provenance architecture tutorial, March 2006. (url: http://www.gridprovenance.org/architecture/tutorial.html).

[383] Curt Tilmes. Provenance tracking in climate science data processing systems. In *Eos Trans. AGU*, volume 89, 2008. abstract IN11C-1042, (url: http://www.agu.org/cgi-bin/wais?mm=IN11C-1042).

[384] Curt Tilmes and Albert J. Fleig. Provenance tracking in an earth science data processing system. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 221–228. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_23).

[385] Paul Townend, Paul Groth, Nik Looker, and Jie Xu. Ft-grid: A fault-tolerance system for e-science. In *Proceedings of the UK OST e-Science Fourth All Hands Meeting (AHM05)*, September 2005. (url: http://www.allhands.org.uk/2005/proceedings/papers/392.pdf).

[386] Paul Townend, Paul Groth, and Jie Xu. A provenance-aware weighted fault tolerance scheme for service-based applications. In *Proc. of the 8th IEEE International Symposium on Object-oriented Real-time distributed Computing (ISORC 2005)*, pages 258–266, Los Alamitos, CA, USA, May 2005. IEEE Computer Society. (doi: http://doi.ieeecomputersociety.org/10.1109/ISORC.2005.3).

[387] W. Tsai, Xiao Wei, Yinong Chen, Ray Paul, Jen-Yao Chung, and Dawei Zhang. Data provenance in soa: security, reliability, and integrity. *Service Oriented Computing and Applications*, 1(4):223–247, December 2007. (doi: http://dx.doi.org/10.1007/s11761-007-0018-8).

[388] Wei-Tek Tsai, Xiao Wei, Dawei Zhang, Ray Paul, Yinong Chen, and Jen-Yao Chung. A new soa data-provenance framework. In *Eighth International Symposium on Autonomous Decentralized Systems. (ISADS'07)*, pages 105–112, March 2007. (doi: http://doi.ieeecomputersociety.org/10.1109/ISADS.2007.5).

[389] Stijn Vansummeren and James Cheney. Recording provenance for sql queries and updates. *IEEE Data Eng. Bull.*, 30(4):29–37, 2007. (url: http://sites.computer.org/debull/A07dec/stijn.pdf).

[390] J. Vázquez-Salceda, S. Alvarez, T. Kifor, L. Z. Varga, S. Miles, L. Moreau, and S. Willmott. *In R. Annicchiarico, U. Cortés, C. Urdiales (eds.) Agent Technology and E-Health*, chapter EU PROVENANCE Project: An Open Provenance Architecture for Distributed Applications, pages 45–63. Whitestein Series in Software Agent Technologies and Autonomic Computing. Birkhauser Verlag AG, Switzerland, December 2007. (doi: http://dx.doi.org/10.1007/978-3-7643-8547-7_4).

[391] Javier Vazquez-Salceda and Sergio Alvarez-Napagao. Using soa provenance to implement norm enforcement in e-institutions. In *Proceedings of the Workshop on Coordination, Organizations, Institutions and Norms (COIN@AAAI08)*, pages 188–203, Berlin, Heidelberg, 2009. Springer-Verlag. (doi: http://dx.doi.org/10.1007/978-3-642-00443-8_13).

[392] Nithya Vijayakumar. *Data Management in Distributed Stream Processing Systems*. PhD thesis, University of Indiana, 2007.

[393] Nithya N. Vijayakumar and Beth Plale. Towards low overhead provenance tracking in near real-time stream filtering. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 46–54. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_6).

[394] Liqiang Wang, Shiyong Lu, Xubo Fei, Artem Chebotko, H. Victoria Bryant, and Jeffrey L. Ram. Atomicity and provenance support for pipelined scientific workflows. *Future Generation Computer Systems*, 25(5):568 – 576, 2009. (doi: http://dx.doi.org/10.1016/j.future.2008.06.007).

[395] Liqiang Wang, Shiyong Lu, Xubo Fei, and Jeffrey Ram. A dataflow-oriented atomicity and provenance system for pipelined scientific workflows. In *Proc. 2nd International Workshop on Workflow Systems in e-Science(WSES 07)*, in conjunction with *International Conference on Computational Science (ICCS) 2007*, volume 4489 of *Lecture Notes in Computer Science*. Springer, 2007. (doi: http://dx.doi.org/10.1007/978-3-540-72588-6_42).

[396] Min Wang, Marion Blount, John Davis, Archan Misra, and Daby Sow. A time-and-value centric provenance model and architecture for medical event streams. In *HealthNet '07: Proceedings of the 1st ACM SIGMOBILE international workshop on Systems and networking support for healthcare and assisted living environments*, pages 95–100, New York, NY, USA, 2007. ACM. (doi: http://doi.acm.org/10.1145/1248054.1248082).

[397] Shaowen Wang, Anand Padmanabhan, James D. Myers, Wenwu Tang, and Yong Liu. Towards provenance-aware geographic information systems. In *GIS '08: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, pages 1–4, New York, NY, USA, 2008. ACM. (doi: http://doi.acm.org/10.1145/1463434.1463515).

[398] Y. Richard Wang and Stuart E. Madnick. A polygon model for heterogeneous database systems: the source tagging perspective. In *Proceedings of*

the sixteenth international conference on Very large databases, pages 519–533, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc. (url: http://web.mit.edu/tdqm/www/tdqmpub/polygenmodelAug90.pdf).

[399] Rowland E. Watkins and Denis A. Nicole. Named graphs as a mechanism for reasoning about provenance. In *Frontiers of WWW Research and Development - APWeb 2006: 8th Asia-Pacific Web Conference*, volume 3841 of *Lecture Notes in Computer Science*, pages 943–948, Harbin, China, 2006. (doi: http://dx.doi.org/10.1007/11610113).

[400] Andrea Weise, Adil Hasan, Mark Hedges, and Jens Jensen. Managing provenance in irods. In Gabrielle Allen, Jaroslaw Nabrzyski, Edward Seidel, G. Dick van Albada, Jack Dongarra, and Peter M. A. Sloot, editors, *Computational Science - ICCS 2009, 9th International Conference, Baton Rouge, LA, USA, May 25-27, 2009, Proceedings, Part II*, volume 5545 of *Lecture Notes in Computer Science*, pages 667–676. Springer, 2009. (doi: http://dx.doi.org/10.1007/978-3-642-01973-9_75).

[401] D. J. Weitzner, H. Abelson, T. Berners-Lee, C. Hanson, J. Hendler, L. Kagal, D. L. McGuinness, G. J. Sussman, and K. K. Waterman. Transparent accountable data mining: New strategies for privacy protection. Technical Report MIT-CSAIL-TR-2006-007, Massachusets Institute of Technology Computer Science and Artificial Intelligence Laboratory, 2006. (url: http://dig.csail.mit.edu/2006/01/tami-privacy-strategies-aaai.pdf).

[402] Daniel J. Weitzner, Harold Abelson, Tim Berners-Lee, Joan Feigenbaum, James Hendler, and Geral Jay Sussman. Information accountability. *Commun. ACM*, 51(6):81–87, June 2008. (doi: http://doi.acm.org/10.1145/1349026.1349043).

[403] J. Widom. Trio: a system for integrated management of data, accuracy, and lineage. In *Second Biennial Conference on Innovative Data Systems Research (CIDR 2005)*, Asilomar, Calif., January 2005. (url: www.cidrdb.org/cidr2005/papers/P22.pdf).

[404] Sylvia C. Wong, Simon Miles, Weijian Fang, Paul Groth, and Luc Moreau. Provenance-based validation of e-science experiments. In *Proceedings of 4th Internation Semantic Web Conference (ISWC'05)*, volume 3729 of *Lecture Notes in Computer Science*, pages 801–815, Galway, Ireland, November 2005. Springer-Verlag. (doi: http://dx.doi.org/10.1007/11574620_57).

[405] Sylvia C. Wong, Simon Miles, Weijian Fang, Paul Groth, and Luc Moreau. Validation of e-science experiments using a provenance-based approach. In *Proceedings of Fourth All Hands Meeting (AHM'05)*, Nottingham, September 2005. (url: http://eprints.ecs.soton.ac.uk/11063/).

[406] A. Woodruff and M. Stonebraker. Supporting fine-grained data lineage in a database visualization environment. In *Proceedings of the 13th International Conference on Data Engineering*, pages 91–102, Birmingham, England, April 1997. IEEE Computer Society. (doi: http://doi.ieeecomputersociety.org/10.110910.1109/ICDE.1997.581742).

[407] Allison Gyle Woodruff. *Data Lineage and Information Density in Database Visualization*. PhD thesis, University of California at Berkeley, 1998. (url: http://db.cs.berkeley.edu/papers/UCB-PhD-woodruff.pdf).

[408] Ian Wootten, Shrija Rajbhandari, Omer Rana, and Jaspreet Pahwa. Actor provenance capture with ganglia. In *Proceedings of the 6th IEEE Internation Symposium on Cluster Computing and the Grid (CCGrid'06)*, pages 99–106, Washington, DC, USA, 2006. IEEE Computer Society. (doi: http://dx.doi.org/10.1109/CCGRID.2006.1).

[409] Ian Wootten, Omer Rana, and Shrija Rajbhandari. Recording actor state in scientific workflows. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 109–117. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_13).

[410] Ian Wootten and Omer F. Rana. Recording the context of action for process documentation. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 45–53. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_6).

[411] Jie Xu, Paul Townend, Nik Looker, and Paul T. Groth. Ft-grid: a system for achieving fault tolerance in grids. *Concurrency and Computation: Practice and Experience*, 20(3):297–309, 2008. (doi: http://dx.doi.org/10.1002/cpe.1266).

[412] Jing Zhang, Adriane Chapman, and Kristen LeFevre. Fine-grained tamper-evident data pedigree. Technical Report CSE-TR-548-08, University of Michigan, 2008. (url: https://www.eecs.umich.edu/eecs/research/techreports/cse_tr/database/reports.cgi?08).

[413] Mingwu Zhang, Daisuke Kihara, and Sunil Prabhakar. Tracing lineage in multi-version scientific databases. In *IEEE 7th International Symposium on Bioinformatics and Bioengineering (BIBE)*, pages 440–447, 2007. (doi: http://dx.doi.org/10.1109/BIBE.2007.4375599).

[414] Mingwu Zhang, Xiangyu Zhang, Xiang Zhang, and Sunil Prabhakar. Tracing lineage beyond relational operators. In *VLDB '07: Proceedings of*

the 33rd international conference on Very large data bases, pages 1116–1127. VLDB Endowment, 2007. (url: http://www.cs.purdue.edu/homes/sunil/pub/BlackBox.pdf).

[415] J. Zhao, C. Goble, M. Greenwood, C. Wroe, and R. Stevens. Annotating, linking and browsing provenance logs for e-science. In *Proceedings of the ISWC 2003 Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data*, Sanibel Island, Florida, October 2003. (url: http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-83/prov_2.pdf).

[416] J. Zhao, C. Goble, R. Stephens, and S. Bechhofer. Semantically linking and browsing provenance logs for e-science. In *Proc. of the 1st International Conference on Semantics of a Networked World*, volume 3226 of *Lecture Notes in Computer Science*, pages 158–176, Paris, France, June 2004. Springer. (doi: http://dx.doi.org/10.1007/b102069).

[417] Jun Zhao. *A conceptual model for e-science provenance*. Ph.d. thesis, University of Manchester, June 2007. (url: http://users.ox.ac.uk/~zool0770/jun_thesis_final_2007.pdf).

[418] Jun Zhao, Carole Goble, and Robert Stevens. An identity crisis in the life sciences. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 254–269. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_26).

[419] Jun Zhao, Carole Goble, Robert Stevens, and Daniele Turi. Mining taverna's semantic web of provenance. *Concurrency and Computation: Practice and Experience*, 20(5):463–472, 2008. (doi: http://dx.doi.org/10.1002/cpe.1231).

[420] Jun Zhao, Alistair Miles, Graham Klyne, and David Shotton. Linked data and provenance in biological data webs. *Brief Bioinform*, pages bbn044+, December 2008. (doi: http://dx.doi.org/10.1093/bib/bbn044).

[421] Jun Zhao, Chris Wroe, Carole Goble, Robert Stevens, Dennis Quan, and Mark Greenwood. Using semantic web technologies for representing e-science provenance. In *Proceedings of Third International Semantic Web Conference (ISWC2004)*, volume 3298 of *Lecture Notes in Computer Science*, pages 92–106, Hiroshima, Japan, November 2004. Springer-Verlag. (doi: http://dx.doi.org/10.1007/b102467).

[422] Yong Zhao. *A Virtual Data Language and System for Scientific Workflow Management in Data Grid Environments*. PhD thesis, THE UNIVERSITY OF CHICAGO, August 2007.

[423] Yong Zhao and Shiyong Lu. A logic programming approach to scientific workflow provenance querying. In Juliana Freire, David Koop, and Luc Moreau, editors, *Second International Provenance and Annotation Workshop, IPAW'2008*, volume 5272 of *Lecture Notes in Computer Science*, pages 31–44. Springer, June 2008. (doi: http://dx.doi.org/10.1007/978-3-540-89965-5_5).

[424] Yong Zhao, Michael Wilde, and Ian Foster. Applying the virtual data provenance model. In Luc Moreau and Ian Foster, editors, *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW'2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 148–161. Springer, 2006. (doi: http://dx.doi.org/10.1007/11890850_16).

[425] Wenchao Zhou, E. Cronin, and Boon Thau Loo. Provenance-aware secure networks. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, pages 188 – 193, April 2008. (doi: http://dx.doi.org/10.1109/ICDEW.2008.4498315).

Bibliography

[426] W3C Incubator Activity. Provenance incubator group charter, September 2009. (url: http://www.w3.org/2005/Incubator/prov/charter).

[427] Scott Bateman, Carl Gutwin, and Miguel Nacenta. Seeing things in the clouds: the effect of visual features on tag cloud selections. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 193–202, New York, NY, USA, 2008. ACM. (doi: http://doi.acm.org/10.1145/1379092.1379130).

[428] Tim Berners-Lee. Linked data. Technical report, World Wide Web Consortium, 2006. (url: http://www.w3.org/DesignIssues/LinkedData.html).

[429] Tim Berners-Lee, Wendy Hall, James A. Hendler, Kieron O'Hara, Nigel Shadbolt, and Daniel J. Weitzner. A framework for web science. *Found. Trends Web Sci.*, 1(1):1–130, 2006. (doi: http://dx.doi.org/10.1561/1800000001).

[430] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, 2001. (url: http://www.scientificamerican.com/article.cfm?id=the-semantic-web).

[431] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009. (url: http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf).

[432] Scott Boag, Don Chamberlin, Mary F. Fernandez, Daniela Florescu, Jonathan Robie, and Jerome Simeon. Xquery 1.0: An xml query language. W3c recommendation, World Wide Web Consortium, January 2007. (url: http://www.w3.org/TR/xuqery/).

[433] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, and Ivona Brandic. Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Gener. Comput. Syst.*, 25(6):599–616, 2009. (doi: http://dx.doi.org/10.1016/j.future.2008.12.001).

[434] Chaomei Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *J. Am. Soc. Inf. Sci. Technol.*, 57(3):359–377, 2006. (doi: http://dx.doi.org/10.1002/asi.v57:3).

[435] Chaomei Chen, Yue Chen, Mark Horowitz, Haiyan Hou, Zeyuan Liu, and Don Pellegrino. Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, 2009. (doi: http://dx.doi.org/10.1016/j.joi.2009.03.004).

[436] James Clark and Steve DeRose. Xml path language (xpath) version 1.0. W3c recommendation, World Wide Web Consortium, November 1999. (url: http://www.w3.org/TR/xpath/).

[437] Mike Dean (ed), Guus Schreiber (ed.) Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. OWL web ontology language refernce. W3c recommendation, World Wide Web Consortium, February 2004. (url: http://www.w3.org/TR/owl-ref/).

[438] L.M. Dusseault, Editor. Http extensions for web distributed authoring and versioning (webdav). Technical report, IETF, June 2007. (url: http://www.webdav.org/specs/rfc4918.html).

[439] Roy Thomas Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2000. (url: http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm).

[440] Ian Foster and Carl Kesselman, editors. *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufman Publishers, 1998.

[441] Ian Foster, Carl Kesselman, and Steve Tuecke. The Anatomy of the Grid. Enabling Scalable Virtual Organizations. *International Journal of Supercomputer Applications*, 15(3):200–222, 2001. (doi: http://dx.doi.org/10.1177/109434200101500302).

[442] Yolanda Gil and Donovan Artz. Towards content trust of web resources. *Web Semant.*, 5(4):227–239, 2007. (doi: http://dx.doi.org/10.1016/j.websem.2007.09.005).

[443] James Hendler. COMMUNICATION: Enhanced: Science and the Semantic Web. *Science*, 299(5606):520–521, 2003. (doi: http://dx.doi.org/10.1126/science.1078874).

[444] A J G Hey and A E Trefethen. The data deluge: An e-science perspective. 2003. (url: http://eprints.ecs.soton.ac.uk/7648/1/The_Data_Deluge.pdf).

[445] Ian Jacobs and Norma Walch. Architecture of the world wide web, volume one. Technical report, World Wide Web Consortium, 2004. (url: http://www.w3.org/TR/webarch/).

[446] Graham Klyne and Jeremy J. Carroll. Resource description framework (rdf): Concepts and abstract syntax. W3c recommendation, World Wide Web Consortium, February 2004. (url: http://www.w3.org/TR/rdf-concepts/).

[447] Luc Moreau. Provenance architecture principles according to rest guidelines. Technical report, University of Southampton, 2009. In Preparation.

[448] Mike P. Papazoglou and Willem-Jan van den Heuvel. Service oriented architectures: Approaches, technologies and research issues. *VLDB Journal*, 16(3):389–415, 2007. (url: http://dx.doi.org/10.1007/s00778-007-0044-3).

[449] Andy Powell, Mikael Nilsson, Ambjorn Naeve, Pete Johnston, and Thomas Baker. Dcmi abstract model. Dcmi recommendation, Dublin Core Metadata Initiative, June 2007. (url: http://dublincore.org/documents/abstract-model/).

[450] Eric Prud'hommeaux and Andy Seaborne. Sparql query language for rdf. W3c recommendation, World Wide Web Consortium, 2008. (url: http://www.w3.org/TR/rdf-sparql-query/).

[451] Sarvapali D. Ramchurn, Dong Huynh, and Nicholas R. Jennings. Trust in multi-agent systems. *Knowl. Eng. Rev.*, 19(1):1–25, 2004. (doi: http://dx.doi.org/10.1017/S0269888904000116).

[452] Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101, 2006. (doi: http://doi.ieeecomputersociety.org/10.1109/MIS.2006.62).

[453] W. T. Luke Teacy, Jigar Patel, Nicholas R. Jennings, and Michael Luck. Coping with inaccurate reputation sources: experimental analysis of a probabilistic trust model. In *AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 997–1004, New York, NY, USA, 2005. ACM. (doi: http://doi.acm.org/10.1145/1082473.1082624).