D-Grid-Integrationsprojekt 2

Diskussionspapier

A: Status des Dokuments
Version 1.0, Draft.

1 The Foundations for Provenance on the Web

Ein Übersichtsartikel von Luc Mureau
(http://eprints.ecs.soton.ac.uk/21691/1/survey.pdf)
The Foundations for Provenance on the Web Luc Moreau, November 11, 2010 (Foundations and Trends in Web Science Vol. 2, Nos. 2-3 (2010) 99–241, L. Moreau, DOI: 10.1561/1800000010)

As the e-science vision becomes reality, 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, or databases resulting from sophisticated curation. While science is becoming computation and data intensive, the fundamental tenet of the scientific method remains unchanged: experimental results need to be reproducible.

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 . . . has the potential to make systems more transparent, and therefore auditable. 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.

Provenance, as a technical subject of study, is by no means a green field. The oldest publications discussed in this survey date back to the late eighties. Importantly, the interest of provenance has been growing dramatically, as illustrated by the number of publications on the topic. Over 400 publications on provenance have been identified, 200 of which have been published over the last two years.

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 a foundational framework is proposed here.

1.1 Analysis of the Provenance Literature

Der Autor untersucht zunächst mittels einer Zitaten-Analyse die Entwicklung der Provenienz-Forschung und versucht ihre Richtung zu bestimmen.

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. Figure 1.1 contains a histogram displaying the number of publications on provenance per year. The bibliography contains papers that were known to the author up to summer 2009. A total of 425 papers have been identified. The first publication dates back from 1986 and describes an auditing technique to assist analysts in understanding and validating data results.

In the case of provenance, it could be conjectured that the development of the Grid as a technology for running scientific applications and the UK e-science programme have been two significant external triggering factors that have caused increasing number of researchers to focus on the provenance problem.

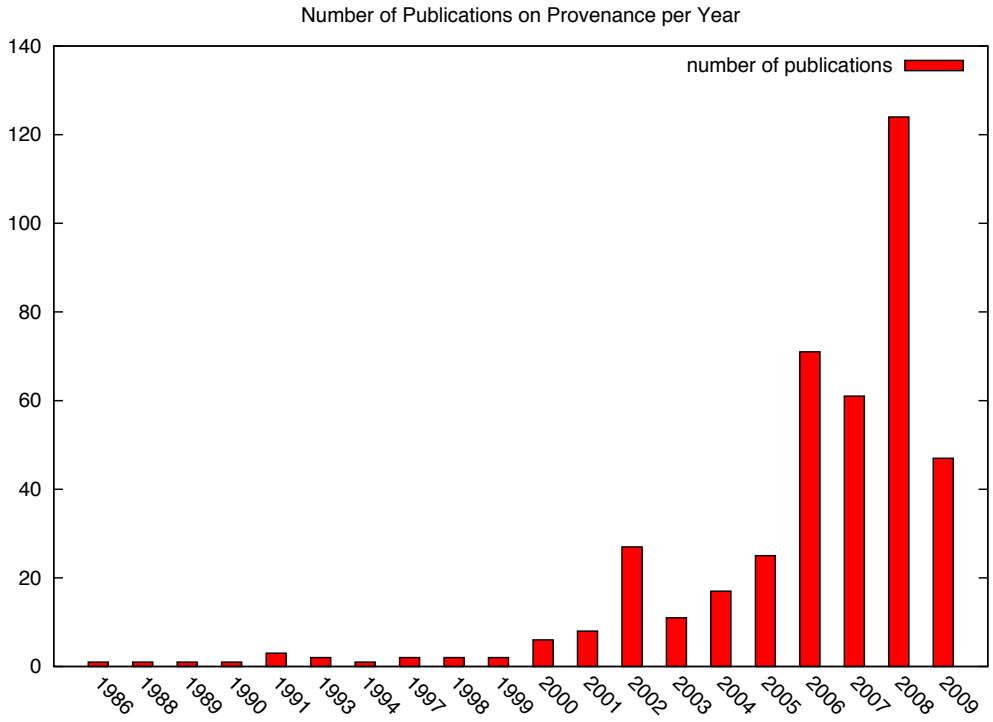

Abbildung 1: Number of Provenance Publications

1.2 Definition of Provenance

Der Autor betrachtet verschiedene Wörterbuch-Definitionen von „Provenance“ und schlägt dann vor:

Definition 1.1. (Provenance as Process) The provenance of a piece of data is the process that led to that piece of data.

Definition 1.1 is concerned with provenance as a concept since potentially many things pertaining to execution may be captured under „process“, including the executed program, input data, configuration, computer, electricitiy powering it, users, etc. From a Computer Science perspective, the goal is to conceive a computer-based representation of provenance that permits useful analysis and reasoning.

Der Autor betrachtet dann weitere Definitionen bzw. Herangehensweisen an Herkunftsinformationen:

- Provenance as Process
- Provenance as a Directed Acyclic Graph
- Why-Provenance
- Where-Provenance
- How-Provenance
- Provenance as Annotations

– Other definitions

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

- **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.
- **Program** Some approaches assume that both the programming language and the program that executed the process are known, and therefore can be used to identify provenance, to derive a reverse function that computes provenance, or to encode provenance efficiently.Other approaches do not make such an assumption, but instead rely on ontological descriptions of what happened, e.g., Provenir, Web Provenance, OPM, PASOA.
- **Trusted base** For language-aware approaches, the trusted base is the compiler and runtime system compliant with a language definition (e.g. SQL or Java); for ontological-based approaches, the trusted base is the certified library, service or workflow, whose operations are described by ontologies.
- **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.
- **What is in the provenance?** Data items that are the original raw value as it entered a system, data items that were the cause of a given data element, and variants where a summary of how the data were used is incorporated.
- **The provenance of what?** For instance, Michlmayr et al. propose the concept of service provenance. Alternatively, Freire et al. introduce the concept of workflow provenance; workflows are a specific kind of data for which the process of derivation also needs to be tracked.
- **Time** It is generally recognized that a provenance model does not have to include time. While time is not mandatory, it is perceived that it is practical for users to be able to refer to time.

For both social and technical reasons, it is impossible to observe the entire behaviour of arbitrary systems, without their cooperation. Instead, one would have to rely on assertions made by the systems' distributed components, about their local actions and involvement in a computation.

. . . [Provenance] can be regarded 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 (Provenance Aware Service Oriented Architecture) approach makes the distinction between assertions about a process (referred to as p-assertions) and provenance obtained by a query over process assertions.

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 the Universe. The scope can identify the systems in which the tracing back should terminate, or the type of source data the user is interested in identifying.

The distinction between process assertions and queries entails a 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, regarding the provenance of data.

1.3 Provenance in Workflows and Databases

Workflow technology is increasingly considered a rapid experiment development tool, with workflow modifications, frequent runs, and parameter tuning; workflow languages are a mechanism to rapidly glue libraries and services, easily transform data, and rapidly automate computational activities. 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.

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, Tan categorizes such a kind of provenance as workflow and coarse-grained; in contrast, fine-grained provenance provides a detailed explanation of how data is actually derived. However, it is not clear that a partition between coarse and fine grained provenance is the right approach: more or less details about services can be exposed, making them grey boxes.

(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.

(ii) A specific aspect of abstraction is concerned with collections of data.

(iii) If the provenance of everything is to be tracked, consideration should be given to storage requirements.

(iv) The means to actually query provenance need to be provided.

(v) Tracking the evolution of workflows is a special kind of provenance tracking.

(vi) Formal properties of provenance are now emerging.

(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.

The Open Provenance Model 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. They may even be offering conflicting views about a same execution observed by two different observers.

Ich stelle mir „Accounts“ (hier im Sinne von Erzählung, Report, nicht Konto!) als verschiedene Sichten auf denselben Arbeitsablauf vor.

Users very frequently have to deal with collections of data, as opposed to individual data items. 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. Representing the provenance of collections and their members is challenging, given all the potential dimensions of the problem: collection mutability, granularity and efficient representation.

Provenance can become huge: in the public database Gene Ontology, the provenance of a single tuple has been observed to be 10Mb; likewise, a 250Mb database of biological data is associated with 6Gb of provenance. The size of provenance matters; because this is a multi-dimensional challenge, it has to be acknowledged that there is a trade-off between compact representation (reducing recording/upload time), compact storage (reducing storage requirements) and query time.

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. Alternatively, PASOA 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.

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.

1.4 The Open Provenance Vision

By and large, provenance approaches developed in the context of databases and workflows deal with closed systems. 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 is introduced in this chapter.

Capturing provenance on the desktop, in the browser, at the operating system level, or at the user interface inevitably brings ethics concerns: should all user interactions with a computer be captured, to the point that all information can be traceable?

The research community has now gained a fairly good understanding on how to make a single monolithic application provenance-aware, by this it is meant 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. Without loss of generality, the extra information to be captured will be referred to as process assertions. 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 (see Figure **??**). (Vgl. dazu auch **??**.)

When data flows across multiple components, the technique described in [the last paragraph] could be adopted to make each individual component provenance-aware. 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. With the Open Provenance Vision, the provenance from individual systems or components can be expressed, connected in a coherent fashion, and queried seamlessly.

To be uniformly queriable, provenance must be represented using ontological descriptions of what happened, so as to be execution technology independent; several such representations are emerging, e.g., Provenir, Web Provenance, OPM, PASOA, which are technology independent. 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. (Siehe dazu auch **??**.)

The Open Provenance Vision postulates that all systems/components should be able to:

1 keep a record of provenance for any important data they produce (in their formats and repositories of choice);

2 follow conventions when exchanging data so that provenance can be traced across systems;

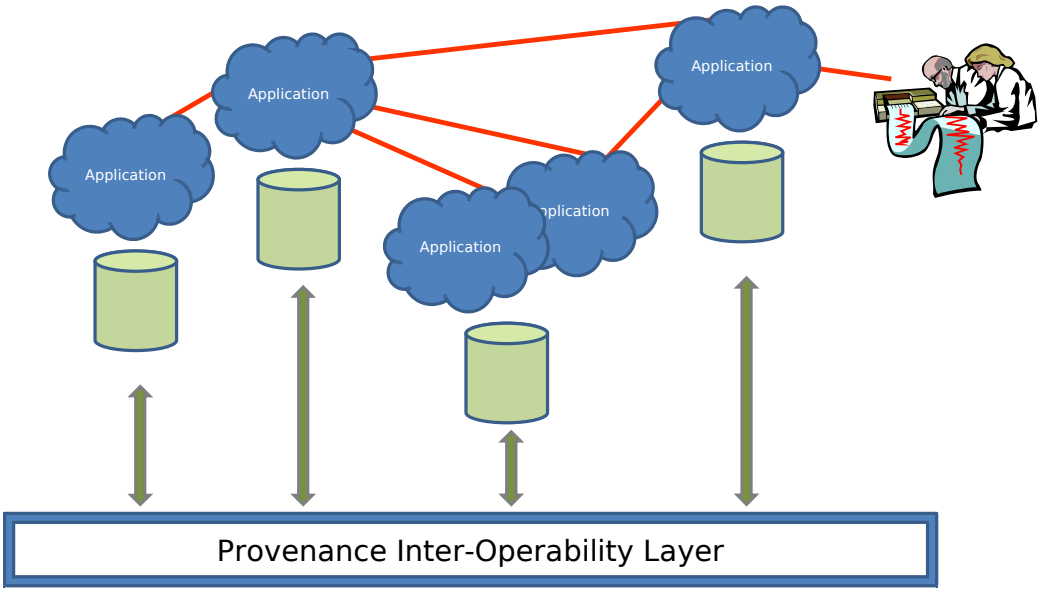

Abbildung 2: The Open Provenance Model (OPM)

3 export provenance of such data using a common data model, such as the Open Provenance Model;

4 answer provenance queries, structured over the common data model.

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. [F]rom 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 computational activities 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.

The Open Provenance Model was designed to represent the provenance of artifacts produced in open systems, by this it is meant systems whose topology may not be known at design time, and whose components, location and identity may only be discovered at runtime.

1.5 Provenance, the Web and the Semantic Web

The ultimate driver for the Open Provenance Vision [...] is the World Wide Web. Technologies such as mashups, tweets 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...

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.

(ii) Semantic Web technologies are themselves being used, not only to represent provenance information, but also to query and reason over it.

(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.

(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.

The principles of exposing information on the Web are now well understood, 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 views, and Webdav.

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 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.

Myers et al. 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 management system, of which Tupelo is a core constituent.

Whilst many authors advocate the use of Semantic Web technologies to represent and query provenance, Carroll et al. 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.

1.6 Accountability

Complex organisations and systems are typically formed by assembling multiple autonomous entities that agree to cooperate in order to achieve overarching objectives. 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, in which autonomous agents collaborate to deliver composite services, provide a means of exploiting such possibilities.

Weitzner et al. note the similarity between accountability and provenance in scientific experiments. Provenance is 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).

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

Rosenthal et al. 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.

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.

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

Factor et al. 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 digital librarians have to modify the original object to have the ability to make it available 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.

An important consideration in any provenance system is the accuracy or objectivity of the assertions recorded. 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.

While it is important to be sure that provenance has not been tampered with, it is also crucial that provenance is faithful, i.e. consisting of an accurate description of execution. Unfortunately, making such a decision in an open environment is not as straighforward as it seems. An unforgeable proof of a component's name does not mean this component faithfully asserted provenance; provenance certification techniques need to be developed to provide better insurance about the trusted base.

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. 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.

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. 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.

Privacy and accountability are both legitimate goals, but they can be at odds.

1.7 Conclusion

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 the community-driven Open Provenance Model. 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.

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.

Dieser Artikel enthält eine 46-seitige **Provenance Bibliography** mit 461 Einträgen.