A Provenance Fabric to Support the Digital Economy

Peter Edwards
Computing Science
University of Aberdeen
United Kingdom
p.edwards@abdn.ac.uk

Edoardo Pignotti
Computing Science
University of Aberdeen
United Kingdom
e.pignotti@abdn.ac.uk

Richard Reid
Computing Science
University of Aberdeen
United Kingdom
r.reid@abdn.ac.uk

ABSTRACT
To support the interface between individuals and organisations it is essential to make the provenance of physical artefacts, digital artefacts, people, organisations, services, service providers and online communications more transparent. In this abstract we outline the requirements for a holistic approach to provenance in the Digital Economy.

Categories and Subject Descriptors
H.1 [Information Systems]: Models and Principles

General Terms
Theory, Documentation, Management

Keywords
Provenance, Linked Data, Trust

1. INTRODUCTION
In recent years, the Web has drastically altered the way in which information and services are exchanged between individuals and organisations. However, these exchanges are often predicated on little or no information as to the quality, reliability or authenticity of the other party. To illustrate, a cloud computing service may provide limited assurance as to its quality, leading to the failure of business-critical applications. We introduce the term provenance fabric to refer to the descriptions of physical artefacts, digital artefacts, people, organisations, services, service providers and online communications that are woven together through the use of linked data principles [1]. Linked data is about exposing, sharing and connecting information on the Web identified by URIs using RDF as a data model. We argue that such an approach has potential to support creation of information that is seen to be trustworthy, reliable and guaranteed. We are exploring the use of Semantic Web technologies in combination with the Open Provenance Model [2] to realise this vision.

Provenance of data has been identified as an essential step in supporting reliability, discovery and trust of online services. Given the challenges facing the Digital Economy, how can we create a solution which captures the context surrounding human activities on the Web, including communication between individuals, society and business? We require a representational framework which goes beyond simple metadata descriptions of digital artefacts, products and services. Such an approach has to bridge both the digital and physical worlds; it must also provide support for reasoning capabilities, to allow it to be used for more than just descriptive purposes. Based upon our experience over a number of years with several applications, we have identified the following requirements for such a provenance fabric:

1. It should describe and uniquely identify a range of entities: artefacts (digital and physical); processes (services and user-centric activities); organisations; people and social networks.
2. It should incorporate online communication (e.g. instant messaging, blog entries, email).
3. It should situate entities in time and space.
4. It should allow relationships (e.g. causal, social, organisational) to be defined between entities.
5. It should make explicit goals and constraints associated with activities, in order to capture the “why?” aspect of provenance.
6. It should facilitate reasoning about policies; trust and reputation.

It is clearly necessary to provide all entities with a unique ID, in order to ensure that humans and machines can accurately refer to the same user, organisation, document, service, etc. However, it may be less obvious that a provenance fabric must span both the physical and digital worlds. While some activities are fulfilled by the execution of computational services, there are still many human-centred activities (e.g. requirements gathering, service design); similarly, there are both digital and physical artefacts and details of all of these must be seamlessly integrated by any provenance fabric. To ensure that we can capture as much of the context associated with processes, it is also important to situate entities within their temporal and spatial context; knowing when and where a process was enacted, or the temporal and spatial relationships that exist between artefacts is important. As we have seen, such a framework must also recognise that electronic communications are important provenance elements. Understanding the reason why a particular activity has been undertaken in terms of the goals of the individual or why a particular artefact has been used provides further important context. Finally, the provenance fabric
must have more than just a descriptive purpose - tools must be able to reason about the content and structure of provenance information in order, for example, to make inferences about the completeness of the record.

2. REALISING A PROVENANCE FABRIC

We now present a partial implementation of a framework realising the provenance fabric. At the heart of our solution is an OWL representation of the Open Provenance Model. This ontology defines the primary entities of OPM (Artefact, Process and Agent) as well as the causal relationships that link them: used, wasGeneratedBy, wasTriggeredBy and wasControlledBy. OPM is a generic solution and as a result, our framework supports additional domain-specific provenance ontologies that are created by extending the concepts defined in the OPM ontology with domain-specific classes. For example, in a service provision ontology, one might have requirements specification as a type of Artefact and service design as a type of Process. To date our investigations have focussed on an e-Science application and we have therefore developed a number of domain-specific provenance ontologies describing aspects of geography, social simulation and biology. Figure 1 illustrates how a physical research activity (an interview) is documented within the provenance fabric as an opm:Process, and how such an activity causes an opm:Artefact to be generated (the interview notes).

The provenance fabric should not only capture information regarding artefacts and processes, but must be able to situate these alongside people and their associated organisational structures. Friend-of-a-Friend (FOAF) is an established RDF vocabulary for describing people and their social networks and we have opted to utilise this within our framework; a foaf:Profile is thus a subclass of the OPM entity Agent. Organisational structures such as groups or institutions can also be defined, and a user may have relationships to any number of such entities. We argue that it is important to capture all actions that occur within a social network and to situate these alongside artefacts and processes within the provenance graph; we will return to this issue later.

The second requirement for the provenance fabric states the need to situate entities within their spatial and temporal context. Our current implementation supports additional causal relationships (e.g. wasCollectedAt, wasSubmittedAt, wasDepositedAt) between opm:Artefacts and GeoNames Features. GeoNames is an ontology that defines geographical locations within the concept of a Feature (which holds information regarding location names, postcodes, coordinates, etc.). We have begun to explore various ways of supplying and navigating temporal provenance, however to date we have yet to fully develop this aspect.

Earlier we argued that capturing online communication is a crucial requirement for achieving a detailed and complete provenance representation. However, the current OPM specifications support limited information about the relationship between person (opm:Agent) and the processes (opm:Process); as there is no regard for relationships associated with the wider social context. As a result, we have integrated the social networking vocabulary - SIOC - with our provenance framework. The SIOC (Semantically-Interlinked Online Communities) ontology is designed to enable the integration of online community information by providing a model to express user-generated content such as posting a message in a blog or posting a comment. Figure 1 illustrates how social context (e.g. sioc:worksWith) and user-generated content (e.g. sioc:Post) can be integrated into the provenance record. The effect of this is a far more definitive provenance representation of the digital artefact; the result of Lorna Philip’s post may have inspired John Farrington to upload the audio recording of the interview. Alternatively, the lack of a vital piece of supporting information may have altered the level of trust in the interview transcript. Therefore, it is crucial to ensure that the highest level of transparency between artefacts, people, organisations, groups, and social interactions exists within the provenance graph.

To date we have implemented only a partial realisation of the provenance fabric. An important requirement of the fabric is to make goals and constraints associated with processes and related artefacts explicit in order to facilitate the understanding of why a particular activity took place. This issue was previously explored by [3] which presents a solution to capture intent (consisting of constraints and goals) on top of a workflow experiment in order to provide additional information about why an experiment was conducted. The example presented in Figure 1 shows what intent was driving John Farrington while conducting the interview process. Reasoning is another aspect of the provenance fabric that is currently not fully realised in our implementation. We are in the process of developing a reasoning apparatus based on a number of existing engines such as Pellet, Fact++ and TrOWL. This will support reasoning about the provenance record, e.g. using rules to describe policies on provenance documentation, or assessing the integrity of the record. Additionally, we should not ignore issues of scalability surrounding the provenance fabric implementation - as it must be capable of providing efficient query answering and reasoning over sizeable amounts of provenance data.

While this approach was originally developed in an e-Science context, we do believe that it is an important step towards our goal of realising a transparent and extensible provenance solution for the Digital Economy. Our current work is investigating how the provenance fabric concept can be deployed in a range of scenarios including sensor networks and crowd-sourced data applications.

3. ACKNOWLEDGMENTS

The research described here was supported by ESRC (RES-149-25-1075) and by EPSRC (EP/G066051/1).

4. REFERENCES


1http://www.foaf-project.org/
2http://www.sioc-project.org
Figure 1: Illustrating the provenance fabric