Future Generation Computer Systems 27 (2011) 766–774

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

W3P: Building an OPM based provenance model for the Web

Andre Freitas [a,*], Tomas Knap [a,b], Sean O'Riain [a], Edward Curry [a]

[a] Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, Ireland
[b] Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic

ARTICLE INFO

Article history:
Received 16 December 2009
Received in revised form
14 October 2010
Accepted 18 October 2010
Available online 26 October 2010

Keywords:
Provenance
Open provenance model
Linked data
Web

ABSTRACT

The Web is evolving into a complex information space where the unprecedented volume of documents and data will offer to the information consumer a level of information integration and aggregation that has up until now not been possible. Indiscriminate addition of information can, however, come with inherent problems such as the provision of poor quality or fraudulent information. Provenance represents the cornerstone element which will enable information consumers to assess information quality, which will play a fundamental role in the continued evolution of the Web. This paper investigates the characteristics and requirements of provenance on the Web, describing how the Open Provenance Model (OPM) can be used as a foundation for the creation of W3P, a provenance model and ontology designed to meet the core requirements for the Web.

1. Introduction

The Web is emerging as a global information space where both documents and data can be reused, aggregated and interconnected in new and unexpected ways. The advent of Linked Data [1] in recent years brings the potential to expose data on the Web, raising new challenges to information consumers. By applying web principles to data, Linked Data allows users to expose data, which was originally limited to database silos, to the Web, lowering the barriers for data linkage and reuse. Since Linked Data can be aggregated and transformed in large chains of information producers and consumers, it is necessary for end users to be able to decide the quality and the trustworthiness of information at hand. Linked Data catalyzes the existing demand for describing the provenance behind information resources on the Web, which can be used as a basis for the assessment of information quality, improving the contextual information behind the generation, transformation and publishing of information on the Web.

Provenance research has been concentrated in the area of scientific workflows in eScience [2]. Consequently, existing works usually approach provenance under the requirements of scientific workflow systems. This focus is shifted in the context of the Web, where provenance should attend a broader audience. Different communities coexist in the Web space, with different perspectives about information, which ultimately drives the way the information is generated or represented. The Web also brings the potential for unexpected usage of information: a specific piece of information can be reused in a completely different context. Since the Web maximizes visibility of information across different communities, provenance becomes the cornerstone element which can help information consumers to assess the quality of information under their quality perspective.

A user facing the decision to use data for a specific purpose should be able to access a representation of the agents, processes and artifacts behind its production and publication. Since this information can be on the open Web, contextual descriptors (e.g. information timeliness) and conditions of use (e.g. digital rights) associated with the data can provide important additional information to the user. Social provenance [3] can be used to determine the trustworthiness on the entities behind an artifact or in the artifact itself. In the context of the Web, provenance, which in scientific workflows was initially focused on the lineage or historical trail of a resource, starts to move towards a comprehensive and structured description of the history, current state and context of an information resource. In addition, the generic use of provenance for quality assessment and trust, common across different Web communities, becomes the fundamental use case for provenance on the Web. In this paper provenance is analyzed under this perspective.

Different communities also have distinct views of provenance. While some consumers may view the quality of information by focusing on the processes which generated the information, others may focus on information publishing aspects. Common across these communities is the need to assess the quality and trustworthiness of the information [4]. In this context, interoperability across different provenance models is central to the process

[*] Corresponding author.
E-mail addresses: andre.freitas@deri.org (A. Freitas), tomas.knap@deri.org, tomas.knap@mff.cuni.cz (T. Knap), sean.oriain@deri.org (S. O'Riain), ed.curry@deri.org (E. Curry).

0167-739X/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2010.10.010

of creating a provenance model for the Web. The Open Provenance Model (OPM) [5], counting with the engagement of a large community in the provenance space, is a strong candidate for becoming the *de facto* provenance interoperability layer. The importance of maximizing interoperability in the process of mapping provenance on the Web and the momentum already achieved in the design of OPM, guided our decision to design W3P, the provenance model described in this paper, to be highly OPM compatible from its inception.

This paper describes the design of W3P, an OPM based provenance model for the Web. Section 2 details the requirements for a provenance model for the Web. As quality assessment is a central motivation for tracking provenance, a discussion about the quality dimensions for the Web is introduced in Section 2.1. A representative set of generic provenance use cases for the Web are described in Section 2.2. These use cases, together with the quality dimensions and supporting literature is the basis for the definition of a set of core requirements for a provenance model for the Web (Section 2.3). The process of building W3P, its compatibility with OPM and a case study of W3P are described in Section 3. Section 4 covers existing related works in the area of provenance on the Web and Section 5 provides conclusions and the future directions for W3P. This paper concentrates its contributions in the requirements analysis for a provenance model for the Web and in the construction of an OPM based model suitable to these requirements.

2. Requirements for a provenance model for the web

The strategy for building the W3P model is based on the creation of a set of requirements for a provenance model for the Web. These requirements are built considering three types of analyses. In a first moment, considering the centrality of provenance as a tool for enabling quality assessment, we investigate a definition for information quality on the Web. Next, four representative use cases of provenance consumption and publishing on the Web are described. The use cases strongly reflect the focus on quality assessment that drives the design of W3P. Later analysis of the use cases provides support for the requirements. The third analysis covers a literature survey to establish a set of core requirements for the provenance model.

2.1. Information quality on the web

The perception of information quality (term used in the literature interchangeably with data quality) is highly dependent on the fitness for use [6] being relative to the specific task that users have at hand. Information quality is usually described in different works by a series of quality dimensions which represent a set of desirable characteristics for an information resource (see [6] for a survey of the main information quality frameworks). The set of information quality dimensions used in this work were composed by the dimensions described in the works of Wang & Strong [7], Alexander & Tate [8] and the set of most common information quality dimensions taken from the comprehensive survey of Knight & Burn [6]. Wang and Strong [7] cover a domain independent set of quality dimensions, while [6,8] cover quality dimensions for the Web. In this work we revisit these dimensions merging them into a single set of dimensions. A small set of the dimensions were omitted since they were not representative for the problem of information quality assessment on the Web or presented some overlap with other dimensions. The final set of information quality dimensions are listed below

1. *Accuracy/correctness:* Represents the extent to which the information is correct and accurate enough for its primary intended use (present in [6–8]).
2. *Compliance:* Covers the extent to which the processes and methodologies behind the data are compliant with the consumers' processes and methodologies (present in [6,7]).
3. *Completeness:* Covers the sufficiency of information for the information consumer (present in [6,7]).
4. *Consistency:* Covers the consistency of the data representation, its model and format in all of its extent (present in [6,7]).
5. *Interpretability:* Represents the quality of the description/ model behind the data. This dimension also covers the suitability of the units or language on which the data is expressed (present in [6,7]).
6. *Usability:* Represents the extent to which the information is helpful for a specific task. In the context of the Web we complement the definition considering the suitability of use in relation to its primary intended use and potential restrictions on the usage of the data (present in [6,7]).
7. *Reputation:* Represents the entities (organizations, individuals) which recommend or repudiate the data, and the trustworthiness of the entities behind the production of a data artifact (present in [6–8]).
8. *Security:* Covers the security mechanisms which enforce the data integrity (present in [6,7]).
9. *Timeliness:* Represents the extent to which the information is sufficiently up to date (present in [6–8]).
10. *Objectivity:* Represents the extent to which the information is unbiased and impartial (present in [6–8]).
11. *Accessibility:* Represents the extent to which the information is available and easily retrievable (from the Linked Data perspective this dimension can represent the appropriate choice and reuse of vocabularies) (present in [6,7]).
12. *Navigation:* Covers the extent to which the data is easily found and linked (present in [6–8]).
13. *Concise:* Represents the extension to which the information is compactly represented (present in [6,7]).

The definition of a standardized provenance model can strongly impact the effectiveness on which consumers enforce their quality criteria. In addition, provenance allows the transfer of trust from entities behind the information to the information itself. Therefore, the creation of a comprehensive provenance model is a fundamental step towards enabling information quality assessment for the Web.

2.2. Provenance use cases

This section contains typical use cases of trust decision and quality assessment for applications consuming and publishing provenance information on the Web. These scenarios, together with quality dimensions and references in the provenance literature, are used to define the key requirements that should be addressed by W3P. These scenarios were developed to maximize the coverage of the use of provenance for the Web, both on document and data level needs. Each use case concentrates on specific provenance problems, with the overlap between some of their features representing the most common provenance uses. The set of use cases summarizes general application areas and are not intended to be an exhaustive investigation of provenance usage in different domains.

2.2.1. Use case I: data integrity and provenance tracking in aggregation of financial data

Description: A financial analyst is using an application that consumes Linked Data from a large number of distributed Web datasets. The datasets include open, government and partner data in the form of stock markets time series, news, blog posts, government data, demographics, previous analysis, third-party qualitative and quantitative analyses and economic facts. The data is directly referenced in a financial report which provides a summarized overview of the economic context of the previous month.

Provenance use: In the process of building the report the analyst uses provenance to determine the trustworthiness of analysis provided by third-party organizations (each organization is an authoritative expert in a specialized market segment). Provenance is also used to determine the analysts (agents) behind the information, since only analysis generated from expert analysts are used. The publisher of the information and its certificate should be available as provenance information in order to be automatically checked. Any news excerpts should have its associated publisher and time information. Each analysis process behind the generation of a report generates a provenance workflow. Each part of the generated workflow has a set of access restrictions expressed in the generated provenance representation.

2.2.2. Use case II: content aggregation and social provenance in a web mashup

Description: A startup is creating a mashup to organize information available on the Web about cars. The website will cover a wide range of interests including press releases, technical specifications, reviews, maintenance tips, brand monitoring, sales offers, etc. Free information available from third parties (e.g. Wikipedia) or information provided by partners will be embedded in the website, while copyrighted content will be exposed as links. Tweets and blog posts will be used to monitor the buzz behind a brand or a car. The information of the mashup will be made available as Linked Data.

Provenance use: The provenance of tweets and blog posts (author, creation/modification date, publisher) needs to be tracked and will be further used for better filtering of the contents. The readers may be able to support or reject a specific resource and this information should be made available as provenance to other readers and to consumers of the Linked Data made available. Every external content embedded on the website should be explicitly quoted and its source, tracked. Usage terms and licensing of third parties of digital artifacts should be represented together with the provenance information.

2.2.3. Use case III: workflow provenance tracking, interoperability, timeliness and licensing for collaboration on the pharmaceutical industry

Description: Pharmaceutical organizations are using the Web to cooperate in a common project for Drug Discovery. Each member of the consortium has access to its internal, partners and public datasets. There are strong cooperation constraints for each partner and trust, security and privacy are key factors to enable an effective collaboration.

Provenance use: Provenance is used to enforce the domain of the partnerships: organization X can cooperate with organization Y in molecular interactions and can cooperate with organization Z in genomic-protein mapping. Each cooperation agreement has an associated time range and terms of usage associated as provenance information with the data. Each group member trusts a different set of public datasets and the provenance of the sources of the data should always be verified. Due to compliance policies and for re-enactment purposes, provenance of the data should also be tracked on the fine-grained experimental workflow level (including infrastructure information: service, machine used, etc.). In this scenario provenance is an important tool for experimental investigation and the ability to query and navigate through the model plays an important role for extracting research value from the information collected. In addition, different members of the consortium use different scientific workflow systems which will need to consume the provenance information of different systems. Vocabularies used in a specific dataset and the linkage with other datasets can provide essential information about the understandability of the data and should be appropriately described. The trail of historical changes of a data item should be preserved. Each member of the consortium has restricted access to the generated provenance information. Data provided by the consortium for the public use has strong constraints on its usage.

2.2.4. Use case IV: geographical and descriptive provenance information for sensor networks

Description: A European consortium in climate change is using a set of environmental sensors distributed in different countries. The data collected from the sensors is published on the Web. Since the sensor infrastructure is inherited from different organizations and application domains, there is a strong inhomogeneity in the conditions and the quality of the data provided. Environment researchers, the end users of the data aggregated from the sensor mashups, need provenance information to determine the quality of the data.

Provenance use: Provenance is used to track the physical location of the sensor, the sensor type/model, timeliness, owner organization, operating conditions, uncertainties associated with the data and measurement units.

2.3. Requirements for a provenance model for the web

Different works in the literature cover distinct perspectives and features of provenance models [2,7–44]. These works will be used, in conjunction with the set of use cases and quality dimensions, to define a set of core requirements for the creation of a provenance model for the Web. Key works in the process of collecting the list of requirements were the extensive survey of the provenance on the Web [2] and the list of requirements for recording and using provenance in eScience experiments [9]. Below, a list of requirements is provided. The requirements are defined by their incidence in the literature, their existence in available web vocabularies/provenance models, their coverage of the use cases and their coverage of the quality dimensions. The requirements detailed here focus on the design of the formal representation (provenance model) and do not address general infrastructure requirements.

1. *Interoperability*: Maximization of the interoperability with existing provenance models and vocabularies. As OPM emerges as a standard interoperability layer across different provenance representations, interoperability can be achieved for a model by the maximization of its OPM compatibility (covered [2,5,10,11]; use cases I, III; quality dimensions 2, 3, 4, 5, 6, 11).
2. *Extensibility*: Support for the addition of domain specific provenance information (requirement based on the multiplicity of provenance models and applications expressed in [12,13,15–22], use case III, quality dimensions 2, 3, 5, 6, 11, 12).
3. *Well defined relational model/logical model/grounded semantics*: Suitability for a wider audience, ability to map to the conceptual model of users, appropriate level of abstraction and grounded semantics, have a strong impact on the usability of the model (use cases I, II, III, IV; quality dimensions 4, 5, 6).
4. *Fine-grained & coarse-grained provenance information*: Ability to express the description of both fine-grained (e.g. statement level) or coarse-grained (dataset/document level) information resources. The provenance model should be able to describe both types of granularities (covered in [23–26], use case III; quality dimensions 2, 3, 5, 6).

5. *Generality*: Coverage of provenance description demands of different communities over the Web (requirement based on the multiplicity of provenance models and applications [12–22], use cases I, II, III, IV; quality dimensions 5, 6, 11).
6. *Data generation & transformation (workflow) description*: Formalization of the description of the processes behind the generation and transformation of the information. For most use cases it is the core of the provenance description and in some scenarios should be fine-grained enough to allow the reproduction of a workflow (including infrastructure aspects). Dependencies between artifacts and justifications are covered by this category of descriptor (covered in [2,5,9,16,17,25,27,28] among other references, use cases I, III; quality dimensions 2, 3, 5, 6).
7. *Spatiality*: Tracking of the geographic location of the information. Spatial information is important in a set of scenarios including tracking of geospatialized artifacts (such as sensor data) and assessment of geospatial trustworthiness and restrictions. (covered in [18,19,29,30]; use case IV; quality dimension 3).
8. *Temporality*: Assessment of the timeliness of the information. Provenance consumers will need to track the temporal evolution of the information resource. (covered in [31–34], concept present in most of the scientific workflows [2,5,9,16, 17,25,28], use cases I, III, IV; quality dimensions 3, 5, 9).
9. *Contracts, digital rights & licensing*: This requirement covers the formalization of the usage conditions of the published artifact (covered in [31–37], use cases II, III; quality dimensions 2, 3, 6).
10. *Integrity mechanisms*: Availability of descriptors for the integrity mechanisms used for both the information resource and its provenance information. Examples are signatures and encryption descriptors (covered in [2,32,38–40], use cases I, III; quality dimension 8).
11. *Identity warranties*: Availability of mechanisms which can provide identity warranties for the elements in the provenance which support an identity (individuals and organizations). Examples of identity warranties are digital certificates (covered in [32,38–40]; use cases I, III; quality dimension 8).
12. *Content description/annotation*: Availability of content descriptors about information resources (tags, titles, natural language descriptions, justifications) (covered in [20,31,41]; use cases I, II, III; quality dimensions 3, 6).
13. *Change tracking:* Ability to describe changes and versioning of an information resource (covered in [31,42], use case III; quality dimension 3).
14. *Coverage of social provenance*: Ability to model trust/distrust/ support/opposition relationships between entities (individuals and organizations) and information resources (covered in [3,38,43], use cases I, II; quality dimensions 3, 7).
15. *Publishing & ownership*: Represents the information related to the publisher entities and processes and the ownership over the information resource (covered in [31,32,38], use cases I, II, III, IV, quality dimensions 2, 5, 7).
16. *Meta-provenance*: Represents descriptors over provenance data, including provenance annotations and permission control over the provenance model entities (covered in [2,39,44], use cases I, III; quality dimension 8).
17. *Query expressivity*: The representation of the provenance model should allow users to launch expressive queries over the model (covered a large set of different works – Section 4.4 of [2], use cases III; quality dimensions 3, 6, 11).
18. *Navigability*: The provenance model should allow users to navigate through its entities (use case III; quality dimensions 3, 6, 11, 12).

In the following section a provenance model compliant with the outlined requirements is described. This provenance model is based on OPM and extends part of its features to address the requirements for mapping provenance on the Web.

3. The W3P provenance model

The previous section discussed the requirements gathering for a provenance model for the Web. This section discusses the construction of W3P, based on the set of requirements and using OPM for the maximization of its interoperability with different provenance models. In addition the coverage of existing vocabularies which could be reused or mapped to W3P is verified. For this purpose a set of *key provenance concepts*, which represent categories that should be covered in a provenance model, were derived from the requirements. Since W3P is built over Web/Semantic Web standards, the suitability of these standards to the requirements is verified and a final model for W3P is discussed together with its mapping to OPM. A description of the application of W3P is described as a case study.

3.1. Building W3P

W3P is designed to provide a general model for representing provenance information on the Web. The model is represented as an ontology and its classes and properties are designed to be intuitive. W3P also works as an integration ontology, providing the structure to reuse already consolidated vocabularies under the more structured semantics of a provenance model. The model is independent of granularity allowing users to describe the provenance of different web artifacts including data, documents and datasets. The coverage of social provenance is one important feature of the ontology, allowing W3P users to track the reputation of entities and artifacts. Fig. 1 depicts W3P excerpts of the classes and properties of the ontology, showing different perspectives including the artifact centered descriptive perspective (1), the workflow perspective (which is the core of the provenance information) (2) and the social relations between entities in the ontology (3). Additional data properties are listed for each core class in (4).

W3P is built over Web/Semantic Web standards (HTTP, URIs, RDF/RDFS [45], OWL [42], SPARQL [46]). The use of Web/Semantic Web standards allows W3P to address the requirements (reqs. 1–5, 17, 18). *Interoperability* (req. 1) is partially covered with the use of widely accepted representation and querying standards provided by SPARQL. The use of the predicates *owl:equivalentClass*, *owl:equivalentProperty* and *owl:sameAs* can map the equivalence of different classes, properties and individuals impacting on *interoperability* and *extensibility*(req. 2). *Extensibility* is one of the built in strengths of Semantic Web, where schemas can be easily extended and merged. *Well Defined Relational Model/Logical Model/Grounded Semantics* (req. 3) is covered by RDF, RDFS and OWL. The use of URIs as identifiers also provides the basic infrastructure for unambiguously expressing concepts, impacting also on this requirement. *Fine-Grained & Coarse-Grained Provenance Information* (req. 4) can be partially addressed with the deployment of reification, named graphs or dataset level descriptors. Semantic Web models provide an expressive and generic way to create representations of provenance models both under a graph or a description logic perspective (*Generality*, req. 5). SPARQL provides an expressive query language for querying the provenance model covering the *Query Expressivity* requirement (req. 17). The use of de-referentiable URIs (one of the principles of the Linked Data Web) and RDF (a graph representation) allows the coverage of the *Navigability* requirement (req. 18).

From the set of requirements (5–16), a collection of *key provenance concepts* were identified. The key provenance concepts represent broader categories which were used to verify the provenance coverage of the vocabularies available on the Web and were also used to design W3P classes and properties. In this paper the vocabularies analyzed were OPM 1.1, the Friend

Table 1
Table mapping the key provenance concepts to the list of vocabularies analyzed (DCMI, FOAF, CS, CC, voiD, OPM, W3P: classes have the first letter capitalized).

Key provenance concepts	Definition	Coverage of vocabulary elements	Reqs.
Certification authority	The authority that issues an identity warranty for the elements in the provenance model which have an identity.	SWP: Complete. DCTERMS: Poor/Incomplete. **W3P uses SWP to cover this key concept.**	5, 11
Publisher	An individual or organization responsible for publishing an information resource.	DCTERMS: Partial. **W3P: Complete.**	5, 15
Owner	The organization or individual which owns the rights over an information resource.	**W3P: Complete.**	5, 15
Host	The organization that provides the infrastructure for the publication of an information resource.	**W3P: Complete.**	5, 15
Integrity mechanisms	The integrity warranties associated with an information resource (e.g. a digital signature).	SWP: Complete. FOAF: Poor/Incomplete. **W3P uses SWP to cover this key concept.**	5, 10
Temporal information	Explicit temporal information that could be associated with the resource. Expiration, creation, modification datetime, and valid range are examples of temporal descriptors.	OPM: Complete. DCTERMS: Medium/Incomplete. **W3P: Complete.**	5, 8
Spatial information	Explicit spatial information associated with the information resource.	WGS84: Complete FOAF: Poor/Incomplete. DCTERMS: Poor/Incomplete. **W3P uses WGS84 to cover this key concept.**	5, 7
License	Descriptors specifying the rights associated with the usage of the data.	CC: Complete (Coarse-grained licensing information). DCTERMS: Complete (Coarse-grained licensing information). **W3P uses both CC and DCTERMS to cover this key concept.**	5, 9
Descriptors/annotations	Human or machine readable descriptions providing a less constrained detailment over the information resource.	OPM: Generic elements for annotations. voiD: Description elements for datasets. FOAF: Covers a basic set of descriptors for the Web. DCTERMS: Covers a large set of descriptors for Web publishing. **W3P: provide a minimal set of descriptors.**	5, 12
Artifact	Any artifact that is the input or the product of a process. An artifact can be the origin or part of a different artifact.	OPM: Complete coverage of the concept (Abstract representation). DCTERMS: Poor/Incomplete. **W3P: Complete.**	5, 6
Process	An operation associated with the generation and transformation of an artifact.	OPM: Complete coverage of the concept (Abstract representation). **W3P: Complete.**	5, 6
Agent (Creator/modifier)	The organization or individual which creates/modifies/access/ interacts with a process or artifact.	OPM: Complete coverage of the concept (Abstract representation). CS: Poor/Incomplete. FOAF: Poor/Incomplete. DCTERMS: Medium/Incomplete. **W3P: Complete.**	5, 6
Social descriptors	Support/opposition of an individual or organization in relation to an artifact or provenance entity. Can also represent the process of quoting an artifact (indirect support).	SWP: Poor/Incomplete. FOAF: Poor/Incomplete. DCTERMS: Poor/Incomplete. **W3P: Complete.**	5, 14
Vocabularies descriptors	The set of vocabularies that are being used to describe an information resource.	voiD: Complete. **W3P uses voiD to cover this key concept, adding an additional property.**	3, 4, 5, 12
Artifact/collection linkage	Contains information about the relationship among the artifacts/collections of artifacts/datasets.	OPM: Medium/Incomplete. DCTERMS: Medium/Incomplete. voiD: Covers dataset linkage. **W3P uses void for dataset linkage and provides its own description for artifact linkage.**	4, 5, 12
Change tracking	Represents the tracking of the changes on the data.	CS: Medium/Incomplete. DCTERMS: Medium/Incomplete. **W3P uses both CS + DCTERMS to cover this key concept.**	5, 13
Meta-provenance	Represents the annotations over provenance descriptors.	**W3P: Complete.**	5, 16
Infrastructure Descriptors	Represents the fine-grained information about the infrastructure behind the generation, transformation or publication of a data artifact.	**W3P: Complete.**	5, 6

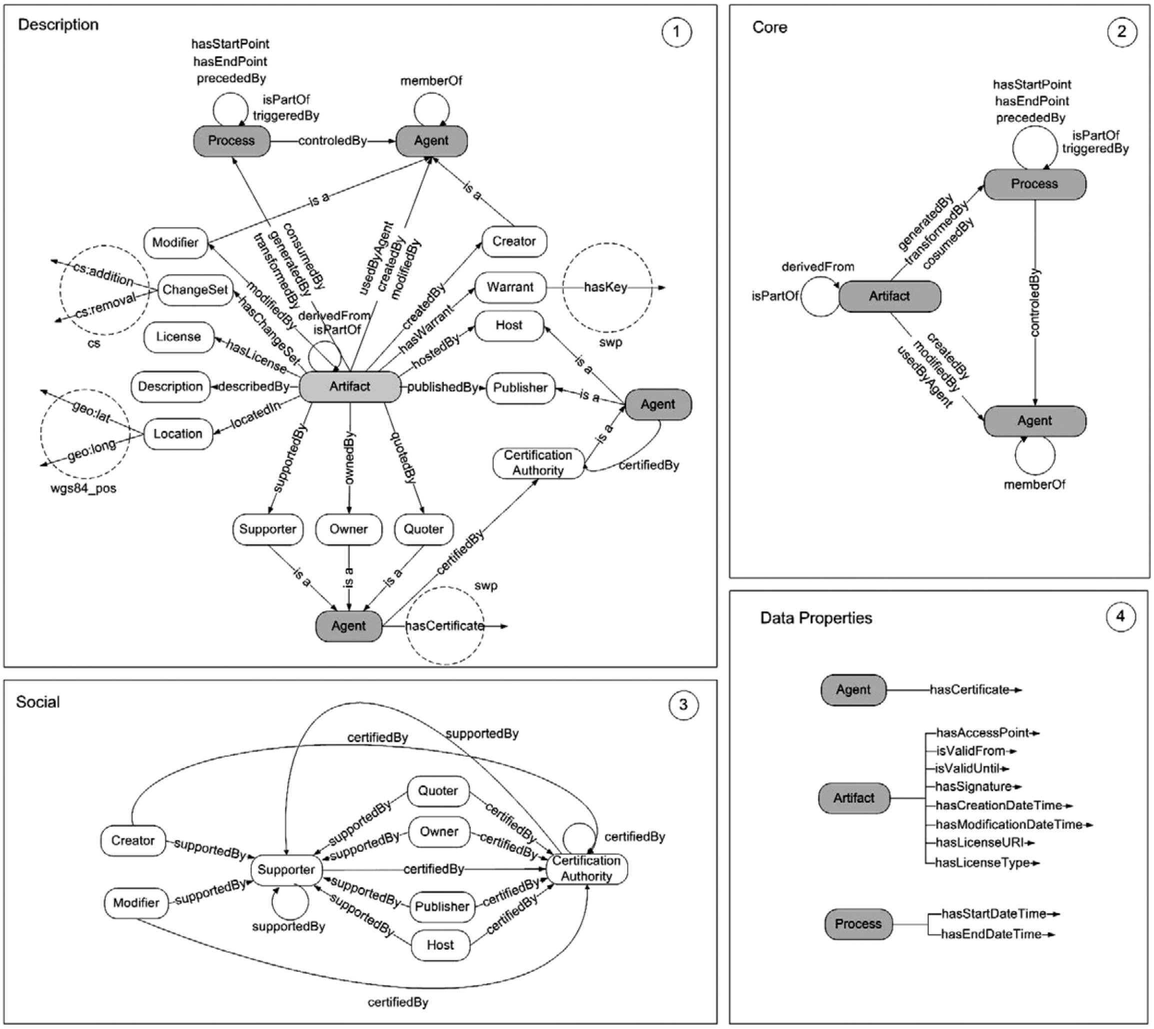


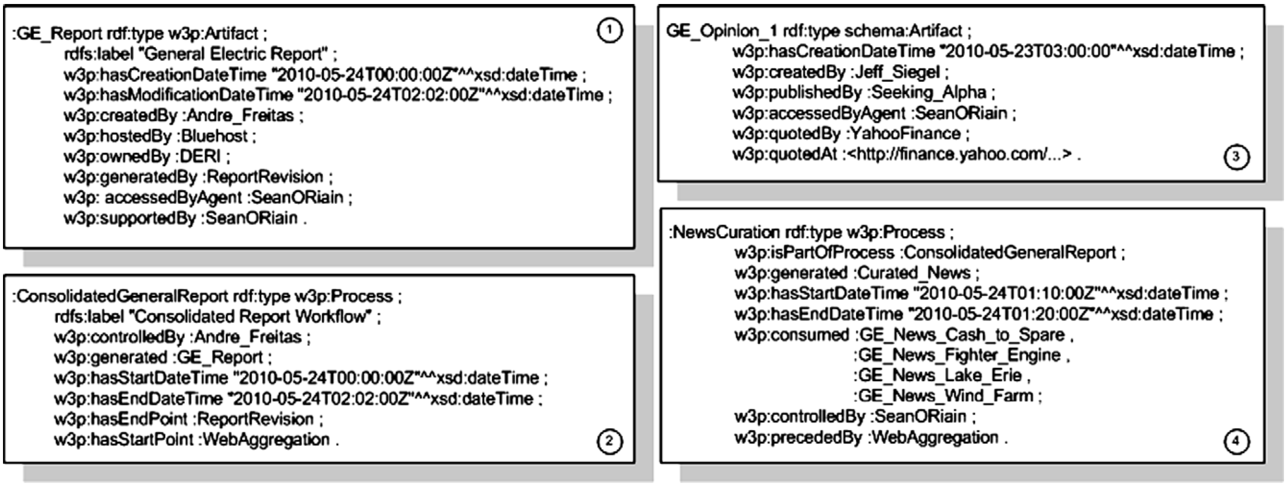
Fig. 1. Partial depiction of W3P.

of a Friend Vocabulary (FOAF) [43], the Dublin Core Metadata Initiative (DCMI) [31], the Semantic Web Publishing Vocabulary (SWP) [38], Creative Commons (CC) [35], WGS84 [29], ChangeSet (CS) [42] and the Vocabulary of Interlinked Datasets (voiD) [41]. The analysis of the existing vocabularies identifies existing gaps in the representation of provenance information and provides the base for the construction of W3P, while maximizing the reuse of existing vocabularies. A summarized analysis of the coverage of the vocabularies and W3P is described in Table 1. A detailed list of the coverage containing the elements present in the vocabularies is described in the W3P documentation.[1]

From Table 1 it is possible to observe that SWP provides good coverage of the *Certifier* and *Integrity Mechanisms* concepts. For the *Spatial Information* concept WGS84 provides a better vocabulary compared DCMI and FOAF. *Publisher*, *Owner* and *Host* are concepts which are not well covered by the analyzed vocabularies. The *Temporal Information* is reasonably well covered by OPM, while DCMI provides a poor set of temporal predicates from the provenance perspective. The key concept of *Social Descriptors* is not well covered by the existing vocabularies: the terms provided by FOAF do not cover the expression of social provenance. *Artifact*, *Process* and *Agent* are well covered by OPM and have a poor coverage in other vocabularies. The CS vocabulary provides a good coverage for the *Change Tracking* key concept (and is complemented by DCMI attributes). The *License* key concept is well covered by both CC and DCMI. For applications which demand a fine-grained model of digital rights/digital contracts these vocabularies are not appropriate. The evolution of initiatives such as the Open Digital Rights Language (ODRL) [47,48] into a vocabulary will cover this missing descriptive gap. *Vocabularies Descriptors* and *Artifact/Collection Linkage* are covered by voiD. DCMI and FOAF provide a comprehensive set of general descriptors which can improve the interpretability of the resource descriptions. The key concept of *Meta-provenance* and *Infrastructure Descriptors* are not covered by other vocabularies.

The dimensions not covered, unstructured or poorly covered from the provenance perspective, defined the scope of W3P. In the process of building the ontology, the reuse of existing vocabularies was maximized. However, reusing concepts which were partially or poorly covered in other vocabularies could lead to a fragmented, inconsistent or difficult to use vocabulary, corrupting the interpretability of the model (req. 3). In addition, some vocabularies were designed to be used as metadata annotations, lacking a more structured model behind them. This is an important design issue which directly impacts requirements 3, 6, 18 (*Well Defined Relational Model/Logical Model/Grounded Semantics, Data Generation & Transformation, Navigability*). OWL primitives for property characteristics (*owl:TransitiveProperty*, *owl:inverseOf*) and ontology mapping (*owl:equivalentClass*, *owl:equivalentProperty*, *owl:sameAs*)

[1] http://prov4j.org/w3p/w3p_coverage.html.



```
:GE_Report rdf:type w3p:Artifact ;
        rdfs:label "General Electric Report" ;
        w3p:hasCreationDateTime "2010-05-24T00:00:00Z"^^xsd:dateTime ;
        w3p:hasModificationDateTime "2010-05-24T02:02:00Z"^^xsd:dateTime ;
        w3p:createdBy :Andre_Freitas ;
        w3p:hostedBy :Bluehost ;
        w3p:ownedBy :DERI ;
        w3p:generatedBy :ReportRevision ;
        w3p: accessedByAgent :SeanORiain ;
        w3p:supportedBy :SeanORiain .
```

```
:ConsolidatedGeneralReport rdf:type w3p:Process ;
       rdfs:label "Consolidated Report Workflow" ;
       w3p:controlledBy :Andre_Freitas ;
       w3p:generated :GE_Report ;
       w3p:hasStartDateTime "2010-05-24T00:00:00Z"^^xsd:dateTime ;
       w3p:hasEndDateTime "2010-05-24T02:02:00Z"^^xsd:dateTime ;
       w3p:hasEndPoint :ReportRevision ;
       w3p:hasStartPoint :WebAggregation .
```

```
GE_Opinion_1 rdf:type schema:Artifact ;
          w3p:hasCreationDateTime "2010-05-23T03:00:00"^^xsd:dateTime ;
          w3p:createdBy :Jeff_Siegel ;
          w3p:publishedBy :Seeking_Alpha ;
          w3p:accessedByAgent :SeanORiain ;
          w3p:quotedBy :YahooFinance ;
          w3p:quotedAt :<http://finance.yahoo.com/...> .
```

```
:NewsCuration rdf:type w3p:Process ;
          w3p:isPartOfProcess :ConsolidatedGeneralReport ;
          w3p:generated :Curated_News ;
          w3p:hasStartDateTime "2010-05-24T01:10:00Z"^^xsd:dateTime ;
          w3p:hasEndDateTime "2010-05-24T01:20:00Z"^^xsd:dateTime ;
          w3p:consumed :GE_News_Cash_to_Spare ,
                       :GE_News_Fighter_Engine ,
                       :GE_News_Lake_Erie ,
                       :GE_News_Wind_Farm ;
          w3p:controlledBy :SeanORiain ;
          w3p:precededBy :WebAggregation .
```

Fig. 2. Excerpt of W3P provenance elements for use case I.

were used. In the context of workflow provenance, transitivity can strongly impact the requirements 3, 6 and 17. Some of the properties defined in W3P have associated inverse properties, which brings flexibility in the usage, bringing better navigability. The complete W3P ontology and its description can be found at http://prov4j.org/w3p/schema#.

3.2. Using OPM for interoperability

OPM provides a solid foundation for modelling workflow provenance (*generation & transformation*) and is the base for W3P interoperability. Directly from OPM 1.1 ontology, the core part of W3P is derived, providing a less abstract workflow model, which is still compatible with OPM. The level of abstraction of the vocabulary behind OPM can bring practical difficulties for an end user of the model, external to the provenance community. The elements of W3P provide a more practical general purpose vocabulary where the three basic elements of OPM, *Artifact, Agent* and *Process* are defined as the W3P basis.

W3P properties *w3p:usedByProcess, w3p:triggeredBy, w3p:generatedBy, w3p:controlledBy* and *w3p:derivedFrom* have their ranges defined to *opm:Used, opm:WasTriggeredBy, opm:WasGeneratedBy, opm:WasControlledBy* and *opm:WasDerivedFrom* respectively. *w3p:createdBy*, *w3p:modifiedBy* and *w3p:usedByAgent* have their ranges mapped to *w3p:Agent*. In W3P, the classes *Quoter, Owner, Host, Publisher, Supporter, OppositionAgent* and *CertificationAuthority* map to *w3p:Agent* which maps to *opm:Agent*. *w3p:Warrant*, *w3p:CertificationAuthority*, *w3p:Location*, *w3p:License*, *w3p:Description* map to *opm:Annotation*. *Access control* and *general annotations* predicates are also mapped to OPM annotations on artifacts. Similar approach was used by Miles to cover descriptive terms in Dublin Core [49]. Currently, *w3p:usedByAgent*, *w3p:createdBy*, *w3p:modifiedBy* are mapped to OPM Processes. A future investigation providing a mapping between W3P elements and OPM using OPM profiles [49] is planned.

The *Account* class present in OPM is mapped to a W3P *Process*. Differently from OPM, a *Process* in W3P is a hierarchical structure composed of other processes. The properties *hasStartPoint*, *hasEndPoint*, *precededBy, succeededBy* and *isPartOf* were introduced in W3P in order to facilitate the creation and navigation over workflows. *precededBy*, *succeededBy* and *isPartOf* are defined as transitive properties and are not mapped to OPM. Inside W3P an *Agent* can have a transitive relationship *memberOf* to a different *Agent*. This type of mechanism is important for the deployment of reputation analysis and access control. Social provenance properties, *w3p:supportedBy, w3p:opposedBy* and *w3p:quotedBy* are mapped as OPM annotations.

3.3. Case study: W3P aggregation of financial data

The W3P ontology was instantiated using the first use case, where different types of financial data collected from distributed external sources are aggregated, curated and analyzed by a team of analysts in order to generate a daily financial report for a specific company. The scenario uses actual financial data from the Web, which is enriched with data from an *aggregation–curation-analysis* workflow. The workflow data is added on the top of the existing aggregated data from the Web using a workflow simulator created for this purpose.

The financial report[2] is composed of different types of data (recommendations, fundamental data, stock data, news, opinions, analysis) which are consolidated and analyzed in the report creation workflow. Each element or collection of elements in the final report has its provenance tracked (*w3p:provenance*). External data which was already aggregated by third parties are represented as source *artifacts* in the report and also have their provenance mappings. In the scenario *social provenance* descriptors play an important role in the process of establishing reputation of external elements among different analysts. Fig. 2 depicts a small excerpt of the provenance descriptor (focusing on W3P elements) for the case study, in n-triples[3] format. (1) covers the descriptor of the final *artifact* (*:GE_Report*); (2) shows the main workflow for the report creation (*:ConsolidatedReportWorkflow*); (3) shows*:GE_Opinion_1*, a descriptor for a blog post (*artifact*) which is used in the financial report and (4) depicts the descriptor of the news aggregation process (*:NewsCuration*). The complete version of the example provenance descriptor for one daily report can be found on the Web.[4]

4. Related work

There is an extensive list of works in the area of provenance models and architectures, focused mainly on the domain of scientific workflows. The reader is directed to [2] for a comprehensive survey in the area. This section covers a short discussion on existing works on the definition of provenance models for the Web.

[2] http://prov4j.org/w3p/scenario/financial.html.

[3] http://www.w3.org/2001/sw/RDFCore/ntriples/.

[4] http://prov4j.org/w3p/scenario/ge_aggregate_provenance_20100524#.

In [32], Hartig proposes a Provenance Model for the Web of Data which generated the *provenance vocabulary*. Hartig proposes two dimensions of provenance for the Web of Data: *data access* and *data creation*. Compared to the *provenance vocabulary*, W3P covers the social dimension of provenance. W3P is also designed to be OPM compatible from the start, maximizing its interoperability. Another fundamental difference is that W3P uses a requirements based approach for its construction. Pinheiro et al. [15] introduced PML (Proof Markup Language), a component of the Inference Web project, which focuses on modelling *knowledge provenance and reasoning information*. The provenance model behind PML, PML-P, focuses on tracking provenance in reasoning systems, where the concept of a proof over inference steps determines the attributes of the provenance model. Due to its own purpose, the attributes of PML-P do not provide coverage of the provenance dimensions for a more comprehensive provenance model for the Web, lacking for example, the dimension of social provenance. Groth et al. [50] describe a generic data model for process documentation (the information that describes a process that has occurred), that allows the answer of provenance questions. The model has a precise conceptual definition and it is evaluated with a mash-up use case from the bioinformatics domain. Both Groth's and this work focus on generic (domain independent) provenance models. A major difference is the approach in the definition of the requirements used in the construction of the models: W3P requirements are targeted towards the coverage of provenance representation and use on the Web, while the model described in [50] approaches the problem through a process documentation perspective. Miles et al. [49] describe a detailed mapping between Dublin Core terms and OPM using OPM profiles, deriving relationships between the two vocabularies. This paper does not explicitly explore the idea of creating OPM profiles for W3P.

The W3C Provenance Incubator Group [51], is defining a comprehensive discussion of use cases and requirements which will provide a roadmap for covering provenance on the Web. The group covers different communities with interests in the provenance space and its final output will become an important guideline for future work on the area. In contrast, the objective of this work is to design a provenance model for the Web, defined over a set of requirements and maximizing the reuse and coverage of existing vocabularies.

5. Conclusion & future work

This paper introduced W3P, a proposed OPM based provenance model for the Web. Since provenance is a key element in the process of quality assessment on the Web, an analysis of a set of key quality dimensions for the Web was introduced. The quality dimensions, together with a set of exemplar use cases and with the support of the literature analysis helped to define the core requirements for a provenance model for the Web. From these requirements, a set of key concepts were derived, providing the base for the classification and analysis of existing vocabularies. These key concepts were used to build W3P, a provenance vocabulary for the Web. W3P is designed to be OPM compatible, responding to the key requirement of interoperability. W3P also covers important key concepts such as social provenance. W3P maximizes the reuse of existing vocabularies, being designed map to existing vocabularies. Semantic Web standards were used to implement the representation of W3P. This paper reinforces the vision that these standards are a suitable way to cope with important requirements for a provenance model.

Future investigations will evolve W3P to a more comprehensive model. The authors intend to provide a comparative analysis between W3P and other web provenance vocabularies (Provenance Vocabulary, PML). A more complete mapping between OPM and W3P using OPM profiles is planned. The coverage of W3P to the requirements raised by the W3C Provenance Incubator Group will be verified. In addition, the suitability of W3P for mapping provenance representations in the context of different scientific workflow systems still needs to be experimentally verified and a description and evaluation of the framework associated with W3P (Prov4J) will be published.

Acknowledgements

The authors thank the reviewers for their highly valuable comments which improved this work in many dimensions. The work presented in this paper has been funded substantially by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

References

[1] T. Berners-Lee, Linked data design issues. Available from: http://www.w3.org/DesignIssues/LinkedData.html.

[2] L. Moreau, The foundations for provenance on the web, Foundations and Trends in Web Science (2009).

[3] A. Harth, A. Polleres, S. Decker, Towards a social provenance model for the web, in: Workshop on Principles of Provenance (PrOPr), 2007.

[4] D. Artz, Y. Gil, A survey of trust in computer science and the Semantic Web, Web Semantics: Science, Services and Agents on the World Wide Web 5 (2) (2007).

[5] L. Moreau, et al., The Open Provenance Model core specification (v1.1), Future Generation Computer Systems (2010).

[6] S.A. Knight, J. Burn, Developing a framework for assessing information quality on the World Wide Web, Informing Science 8 (2005) 159–172.

[7] R. Wang, D. Strong, Beyond accuracy: what data quality means to data consumers, Journal of Management Information Systems 12 (4) (1996) 5–33.

[8] J.E. Alexander, M.A. Tate, Web Wisdom: How to Evaluate and Create Web Page Quality, 1st ed., 1999.

[9] S. Miles, et al., The requirements of recording and using provenance in e-Science experiments, Journal of Grid Computing (2006).

[10] J. Freire, S. Miles, L. Moreau, Second provenance challenge. 2007. Available from: http://twiki.ipaw.info/bin/view/Challenge/.

[11] J. Futrelle, J. Myers, Tracking provenance semantics in heterogeneous execution systems, Concurrency and Computation: Practice and Experience 20 (5) (2008) 555–564.

[12] E.W. Anderson, et al., Provenance in comparative analysis: A study in cosmology, Computing in Science and Engineering 10 (3) (2008).

[13] P. Buneman, et al. A provenance model for manually curated data, in: IPAW.

[14] P. Buneman, W.C. Tan, Provenance in databases, in: SIGMOD Conference.

[15] P.P. da Silva, D.L. McGuinness, R. Fikes, A proof markup language for semantic web services, Information Systems 31 (4) (2006).

[16] S. Davidson, et al. Provenance in scientific workflow systems, in: Data Engineering Bulletin, 2007.

[17] A. Dolgert, et al., Provenance in high-energy physics workflows, Computing in Science and Engineering 10 (3) (2008).

[18] J. Dozier, J. Frew, Computational provenance in hydrologic science: a snow mapping example, Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 367 (1890) 1021–1033.

[19] P.D. Eagan, S.J. Ventura, Enhancing value of environmental data: data lineage reporting, Journal of Environmental Engineering 119 (1) (1993) 5–16.

[20] J. Futrelle, Provenance and annotation of data, in: Provenance and Annotation of Data, Springer, Berlin Heidelberg, 2006, pp. 64–72.

[21] A. Gehani, VEIL: a system for certifying video provenance, in: International Symposium on Multimedia.

[22] Y. Simmhan, B. Plale, D. Gannon, A survey of data provenance in e-science, SIGMOD Record 34 (2005) 31–36.

[23] D.W. Archer, L.M.L. Delcambre, D. Maier, A framework for fine-grained data integration and curation, with provenance, in a dataspace, in: First Workshop on on Theory and Practice of Provenance, 2009.

[24] P. Buneman, S. Khanna, W.C. Tan, Why and where: a characterization of data provenance, in: ICDT.

[25] E. Deelman, et al., Workflows and e-Science: an overview of workflow system features and capabilities, Future Generation Computer Systems 25 (5) (2009) 528–540.

[26] A. Woodruff, M. Stonebraker, Supporting fine-grained data lineage in a database visualization environment, in: ICDE.

[27] S.B. Davidson, J. Freire, Provenance and scientific workflows: challenges and opportunities, in: SIGMOD Conference.

[28] J. Zhao, et al. Using semantic web technologies for representing e-science provenance, in: Third International Semantic Web Conference, ISWC2004, Hiroshima, Japan, Springer-Verlag.

[29] WGS84 geo positioning. Available from: http://www.w3.org/2003/01/geo/wgs84_pos#.

[30] Nicholas, et al., GeoSpatial Semantics. GeoSpatial Semantics. 2007, Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 20–35.

[31] Dublin Core Metadata Initiative. Available from: http://dublincore.org/documents/dcmi-terms/.

[32] O. Hartig, Provenance information in the web of data, in: Proceedings of the Linked Data on the Web (LDOW) Workshop at WWW, Madrid, Spain.

[33] O. Hartig, J. 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) at ISWC, Washington, USA.

[34] L. Gadelha, M. Mattoso, Kairos: an architecture for securing authorship and temporal information of provenance data in grid-enabled workflow management systems, in: IEEE International Conference on eScience.

[35] Creative commons. Available from: http://creativecommons.org/ns#.

[36] R. Garcia, et al. Formalising ODRL semantics using web ontologies, in: Open Digital Rights Language Workshop ODRL05, 2005, ADETTI.

[37] S. Miles, P. Groth, M. Luck, Handling mitigating circumstances for electronic contracts, in: Proceedings of the AISB 2008 Symposium on Behaviour Regulation in Multiagent Systems: The Society for the Study of Artificial Intelligence and Simulation of Behaviour.

[38] C. Bizer, Semantic Web Publishing Vocabulary (SWP) User Manual, 2006.

[39] U. Braun, A. Shinnar, M. Seltzer, Securing provenance, in: Proceedings of the 3rd conference on Hot topics in security, 2008.

[40] J.J. Carroll, et al. Named graphs, provenance and trust, in: International World Wide Web Conference, 2005.

[41] K. Alexander, et al. Describing linked datasets, in: Proceedings of the Linked Data on the Web Workshop, LDOW 2009.

[42] S. Tunnicliffe, I. Davis, Changeset vocabulary. Available from: http://vocab.org/changeset/schema.rdf, 2005.

[43] The Friend of a Friend (FOAF) project. Available from: http://xmlns.com/foaf/spec/.

[44] A. Syalim, Y. Hori, K. Sakurai, Grouping Provenance Information to Improve Efficiency of Access Control, in: Lecture Notes In Computer Science, vol. 5576, 2009.

[45] W.C. Tan, Provenance in databases: past, current, and future, IEEE Data Engineering Bulletin 30 (2007) 3–12.

[46] SPARQL Query Language for RDF. 2008. Available from: http://www.w3.org/TR/rdf-sparql-query/.

[47] R. Garcia, et al. Formalising ODRL semantics using web ontologies, in: Open Digital Rights Language Workshop ODRL05, ADETTI.

[48] L. Wang, et al., Atomicity and provenance support for pipelined scientific workflows, Future Generation Computer Systems 25 (5) (2009) 568–576.

[49] S. Miles, L. Moreau, J. Futrelle, OPM Profile for Dublin Core Terms, 2009.

[50] P. Groth, S. Miles, L. Moreau, A model of process documentation to determine provenance in mash-ups, ACM Transactions on Internet Technology 9 (1) (2009) 1–31.

[51] W3C provenance incubator group. Available from: http://www.w3.org/2005/Incubator/prov/.

Andre Freitas is a graduate student at the eBusiness Unit at the Digital Enterprise Research Institute (NUI Galway). His main research interest areas include Quality Assessment, Provenance and Semantic Search. Before joining DERI Andre worked as product leader, project manager and software engineer in the Semantic Web space and in different industries including Oil & Gas Exploration, IT Security, Medical, Healthcare, Banking and Telecom industries.

Tomas Knap is a PhD student in the Department of Software Engineering at Charles University in Prague, currently working at Digital Enterprise Research Institute (NUI Galway) as a researcher of the *eBusiness and Financial Services Domain (DEB)* unit. His research focuses on Trust architecture for Linked Data environments. Tomas received MSc in Software Engineering from Charles University in Prague.

Sean O'Riain has a background in information technology across research, telecom and IT sectors. Having worked for Hewlett-Packard's European Software Centre, Sean has had considerable industrial experience in large scale data integration, data analysis and database technologies. As part of HP's Semantic Infrastructure Research Group he managed the exploitation and dissemination of semantic technology, including their approaches and tools to enhance technical development and business return across corporate development groups. Now working for DERI he is involved with joint industrial–academic collaborative research programmes, and industrial sponsored research that have investigated semantically enhanced information integration, business analytics and search.

Edward Curry, as a Research Leader within DERI, heads the Translational Research Unit which studies the uptake and impact of new and emerging technologies within industry. His main research interests include linked data integration, data analytics, semantic search, self-* and nature-inspired middleware. He has studied the utilization of these advanced technologies within numerous fortune 500 companies within the Pharmaceutical, Oil & Gas, Financial, Advertising, Manufacturing, Health Care, and Automotive sectors. Edward has worked extensively with industry advising on the adoption patterns, practicalities, and benefits of new technologies to enhance information architectures and flows within their organizations.