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