Future Generation Computer Systems 34 (2014) 1–16

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Implementing interoperable provenance in biomedical research

V. Curcin [a,*], S. Miles [b], R. Danger [a], Y. Chen [b], R. Bache [b], A. Taweel [b]

[a] Department of Computing, Imperial College London, London SW7 2AZ, United Kingdom
[b] Department of Informatics, Kings College London, Strand, London WC2R 2LS, United Kingdom

HIGHLIGHTS

- Provenance provides evidence for validating biomedical research.
- It achieves model-level interoperability of heterogeneous software.
- Implementation can be challenging for teams lacking provenance expertise.
- We present twenty key recommendations to future implementors.
- Work is based on our experiences in two large biomedical projects.

ARTICLE INFO

Article history:
Received 12 April 2013
Received in revised form
7 October 2013
Accepted 3 December 2013
Available online 16 December 2013

Keywords:
Provenance
Biomedical informatics

ABSTRACT

The provenance of a piece of data refers to knowledge about its origin, in terms of the entities and actors involved in its creation, e.g. data sources used, operations carried out on them, and users enacting those operations. Provenance is used to better understand the data and the context of its production, and to assess its reliability, by asserting whether correct procedures were followed. Providing evidence for validating research is of particular importance in the biomedical domain, where the strength of the results depends on the data sources and processes used. In recent times, previously manual processes have become fully or semi-automated, e.g. clinical trial recruitment, epidemiological studies, diagnosis making. The latter is typically achieved through interactions of heterogeneous software systems in multiple settings (hospitals, clinics, academic and industrial research organisations). Provenance traces of these software need to be integrated in a consistent and meaningful manner, but since these software systems rarely share a common platform, the provenance interoperability between them has to be achieved on the level of conceptual models. It is a non-trivial matter to determine where to start in making a biomedical software system provenance-aware. In this paper, we specify recommendations to developers on how to approach provenance modelling, capture, security, storage and querying, based on our experiences with two large-scale biomedical research projects: Translational Research and Patient Safety in Europe (TRANSFoRm) and Electronic Health Records for Clinical Research (EHR4CR). While illustrated with concrete issues encountered, the recommendations are of a sufficiently high level so as to be reusable across the biomedical domain.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Provenance aims to capture the origin of some data through details of the actions and actors involved in its creation. In scientific applications, provenance helps us to understand research results [1]. For instance, a published clinical study may contain a table showing the statistical significance of some treatment on the case group, as opposed to the control sample. Provenance of that table would consist of the statistical algorithms used, their parameterisation, data cleaning that was applied, case and control definitions, and information about the data provider or the data gathering process used. In some circumstances, a part of the process may change over time (e.g. tweaking the case definition), causing the result to change, and the provenance trace can provide clear information about how the result was obtained and how it may be repeated or improved.

Provenance is directly contributing to several important goals that research methodologies are trying to attain. In itself, provenance traces make the research process *auditable*, by providing a standardised account of actions that unfolded during the process execution. Combined with a formal model, such as a business

* Corresponding author. Tel.: +44 7780608682.
E-mail addresses: vasa.curcin@imperial.ac.uk (V. Curcin), simon.miles@kcl.ac.uk (S. Miles), r.danger@imperial.ac.uk (R. Danger), yuhui.chen@kcl.ac.uk (Y. Chen), richard.bache@kcl.ac.uk (R. Bache), adel.taweel@kcl.ac.uk (A. Taweel).

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.future.2013.12.001

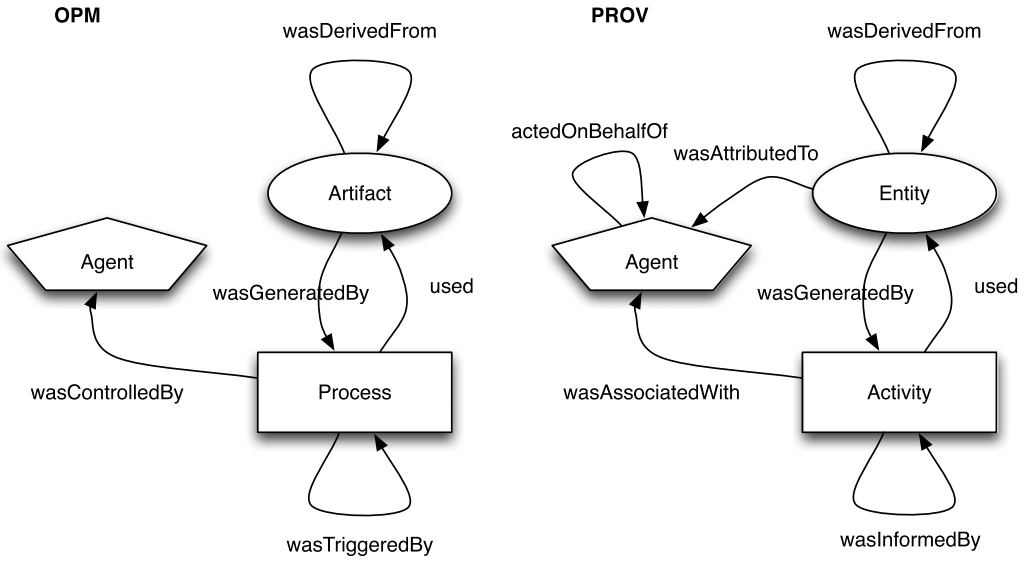

Fig. 1. OPM and PROV type graphs showing available kinds of nodes and edges in each language.

workflow specification, provenance ensures the results are *verifiable*. Finally, when the program executables are provided together with the data used, they jointly ensure *reproducibility* of the research.

Biomedical research is characterised by the heterogeneity of the research teams participating in projects, procedures they follow, and the data they produce. A drug development pipeline would span a range of disciplines from target identification via detection of candidate genes for drugs, to clinical studies exploring the efficacy and drug safety. The need to capture details of data produced at each step and the processes involved is persistent throughout this process and benefits from common technical frameworks that span different scientific domains and multiple teams. For example, collaborative workflows [2] have proven to be highly useful in integrating microarray analysis with low-level gene annotation.

Auditability and verifiability of research data are also essential components of data management in clinical research, due to the sensitivity and importance of its impact on saving lives. This is reflected in popular standards such as GxP (including Good Clinical Data Management Practice and Good Clinical Practice) [3], CONSORT for trial reporting [4], and STROBE [5] for reporting observational studies. Of particular interest is ADAM [6], produced by the Clinical Data Interchange Standards Consortium (CDISC), which documents each derived variable (treatment, outcome, or covariate) used in clinical trial analysis datasets, to enable review and re-creation of published research. All of these standards take a retrospective view of data provenance, as something that needs to be collected and described post-hoc. As will be shown, successful provenance implementations adopt a prospective view, automatically collecting this information in a single repository during the life of a research project.

Reproducibility is also the focus of The Open Data initiative [7], which aims to make publicly generated data free and available to everyone, in useful formats, subject to proper attribution. Another part of that vision, directly relevant to health data management, is that any published research study should be accompanied by the full data that it was derived from, thus enabling the reader to verify the results for themselves. This approach is increasingly taken up by scientific journals [8].

In this paper, we review the implications of provenance for biomedical research by analysing the provenance requirements of two real-world use cases from the clinical research domain, and propose recommendations on appropriate solutions for developing provenance capacity. In particular, we chose use cases that rely on the service oriented architecture paradigm to highlight the importance of provenance in a complex computing system and reveal the benefits that the provenance capacity may bring.

2. Background

The concept of provenance is well established in many disciplines [9,10]. For example, in the study of fine art it refers to the trusted, documented history of some work of art. Electronic tracking of provenance was originally studied in individual domains, including geography or library studies, or with regards to particular technologies, such as databases or workflow systems. It was recognised that the same issues occurred in these different applications, and so similar solutions may apply. Simultaneously, there was a push from many organisations and projects for a standard approach to representing provenance, as this would then allow systems to be developed with some guarantee that the provenance data held would be interpretable in the future. Furthermore, it was understood that an important effect of having common provenance representations would be that the history of data could be traced across multiple heterogeneous systems, as the provenance each system recorded would be interoperable and interconnectable with that recorded by the others.

2.1. Provenance representation models

In the early days of the provenance efforts, several generic provenance models were proposed [11–13]. Several metadata vocabularies also allowed some limited provenance information to be expressed, particularly Dublin Core [14] or Minimum Information About a Microarray Experiment (MIAME) [15] for gene expression data.

Through merging the pioneering efforts of several research groups, a community-driven provenance specification, the Open Provenance Model (OPM) [16] became a de-facto standard representation for provenance in many areas. OPM is a causal graph model, with edges denoting relationships (X was caused by Y) and nodes representing the individual occurrences of entities. The OPM type graph is shown on the left side of Fig. 1. OPM graphs describe the full lineage of a piece of data in terms of multiple events (process instances) that led to it being produced. The nodes in the graph can be of three types: artefacts, processes, and agents.

Download English Version:

https://daneshyari.com/en/article/425911

Download Persian Version:

https://daneshyari.com/article/425911

Daneshyari.com