Journal of Biomedical Informatics 65 (2017) 1–21

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Templates as a method for implementing data provenance in decision support systems

Vasa Curcin [a,*], Elliot Fairweather [a], Roxana Danger [b], Derek Corrigan [c]

[a] Division of Health and Social Care Research, King's College London, London, United Kingdom
[b] Reed Online Ltd, London, United Kingdom
[c] Royal College of Surgeons in Ireland, Dublin, Ireland

ARTICLE INFO

Article history:
Received 15 June 2016
Revised 19 October 2016
Accepted 31 October 2016
Available online 14 November 2016

Keywords:
D2.1 (Software Engineering) Requirements/specification J.3 (Life and Medical Sciences): Health data provenance
Model-driven architectures
Decision support systems

ABSTRACT

Decision support systems are used as a method of promoting consistent guideline-based diagnosis supporting clinical reasoning at point of care. However, despite the availability of numerous commercial products, the wider acceptance of these systems has been hampered by concerns about diagnostic performance and a perceived lack of transparency in the process of generating clinical recommendations. This resonates with the Learning Health System paradigm that promotes data-driven medicine relying on routine data capture and transformation, which also stresses the need for trust in an evidence-based system. Data provenance is a way of automatically capturing the trace of a research task and its resulting data, thereby facilitating trust and the principles of reproducible research. While computational domains have started to embrace this technology through provenance-enabled execution middlewares, traditionally non-computational disciplines, such as medical research, that do not rely on a single software platform, are still struggling with its adoption. In order to address these issues, we introduce provenance templates – abstract provenance fragments representing meaningful domain actions. Templates can be used to generate a model-driven service interface for domain software tools to routinely capture the provenance of their data and tasks. This paper specifies the requirements for a Decision Support tool based on the Learning Health System, introduces the theoretical model for provenance templates and demonstrates the resulting architecture. Our methods were tested and validated on the provenance infrastructure for a Diagnostic Decision Support System that was developed as part of the EU FP7 TRANSFoRm project.

1. Introduction

The importance of data, its origins and quality, has long been recognised in clinical research. In recent years, we have also witnessed increased reliance of clinical practice on data, through routine data capture in Electronic Health Record systems, quality improvement initiatives at multiple levels, and growing adoption of evidence-based medicine.

The patient safety implications of diagnostic error in family practice are potentially severe for both patient and clinician [1]. The development of diagnostic clinical decision support systems (DSS) has long been advocated to promote consistent guideline-based diagnosis supporting clinical reasoning at point of care. However, the wider acceptance of these systems in clinical practice has been much slower in happening despite the availability of many commercial products. Concerns remain about diagnostic performance and a perceived lack of transparency in first generation systems that deploy an evidence knowledge base in the form of a black box that generates clinical recommendations. These concerns about the quality of evidence and the effort required in the longer term maintenance and sustainability of the underlying evidence base supporting such systems has lead to research into second generation tools supporting a more dynamic and iterative cycle of evidence creation and update using a technical infrastructure developed under the auspice of the Learning Health System (LHS) [2].

The Learning Health System community envisages every participant in the health system (clinician, patient, researcher, insurer...) as both a producer and consumer of data. Central to this vision is the notion of routine capture, transformation and dissemination of both data and resulting knowledge. Clinical studies, quality improvement initiatives, decision support, and other scenarios can all then be associated with the routes that the data is taking

* Corresponding author.
E-mail addresses: vasa.curcin@kcl.ac.uk (V. Curcin), elliot.fairweather@kcl.ac.uk (E. Fairweather), roxana.danger@reedonline.co.uk (R. Danger), derekcorrigan@rcsi.ie (D. Corrigan).

http://dx.doi.org/10.1016/j.jbi.2016.10.022
1532-0464/

through the LHS. The trust information associated with the data needs to be made available at each step of these use cases, to support auditability and transparency.

When applied to DSS-s, this trust requirement translates to the ability to readily demonstrate the clinical reasoning that was performed in a clinical encounter, together with the recommendation received. In addition to supporting the auditability of the process, this capability also promotes transparency and traceability from the recommendation back to the rules applied to produce the recommendation. The data provenance community has been working on methods for ensuring reproducibility in scientific research, through use of Semantic Web techniques and the W3C PROV standard [3], that are highly relevant to the challenges of decision support in the LHS environment. Computational provenance provides a uniform data-centred audit trail of what actually happened during some task, and we shall describe how these methods can be adapted to the needs of LHS.

There are two main technical challenges to be addressed in applying data provenance to the Decision Support System scenario; firstly, how to have heterogeneous, distributed software agents (security systems, rule engines...) construct unified, verifiable provenance traces, and secondly, how to formally guarantee that the resulting provenance traces will satisfy domain constraints, often expressed in ontologies, and user data requirements.

In order to address these issues, we introduce *provenance templates*, abstract provenance fragments representing meaningful domain actions that can be used to generate a model-driven service interface for domain software tools to routinely capture the provenance of their data and tasks. A template defines a provenance graph in a generic manner by means of variables such that it may be later instantiated and grafted onto pre-existing provenance graphs. Importantly, this paper introduces the idea that templates may describe subgraphs subject to bounded iteration in both serial and parallel manner.

The EU FP7 TRANSFoRm project [4] has developed a diagnostic decision support tool that promotes numerous state-of-the-art practices of good clinical decision support. These include precisely defined usability patterns, integration with an electronic health record (EHR), allowing for recommendations at the point of care as part of the clinician workflow, and a provenance backend that captures provenance data about the computational aspects of the diagnostic task.

The paper first introduces the concepts of the Learning Health System, data provenance and decision support systems in Section 2, before presenting the requirements of the LHS-enabled DSS, novel provenance templates formalism and the associated provenance architecture in Section 3. Section 4 demonstrates how the new model was used to construct DSS audit trails in TRANSFoRm and in Section 5 we consider how our approach addresses the wider LHS requirements for trust in decision support systems, its impact with respect to some recent developments, and list related work. Section 6 offers conclusions and presents pointers for future research.

2. Background

We shall now review the Learning Health System paradigm and the data provenance technologies, and relate them to the challenges of clinical Decision Support Systems, presenting as an example the DSS developed as part of the TRANSFoRm project.

2.1. Learning Health System

The Learning Health System (LHS) movement aims to establish a next-generation healthcare system, "... one in which progress in science, informatics, and care culture align to generate new knowledge as an ongoing, natural by-product of the care experience, and seamlessly refine and deliver best practices for continuous improvement in health and health care." [5] Each participant in the LHS, be they clinician, patient, or researcher, acts as a consumer and a producer of knowledge, with the LHS providing: (a) routine and secure aggregation of data from multiple sources, (b) conversion of data to knowledge and (c) dissemination of that knowledge, in actionable form, to everyone who can benefit from it [2]. Thus, the LHS creates routes for knowledge transfer between different parts of the health system, thereby increasing its research and learning capacity.

Different data-driven scenarios, such as decision support systems, clinical trial recruitment and management, epidemiological studies, all represent applications within the LHS, each associated with the movements and processing of data and knowledge. A number of LHS implementations have been developed at varying scales [4,6–8].

Attempts to define the core requirements of the Learning Health System [5] have highlighted concerns about a perceived lack of transparency and tracking in current systems demonstrating how clinical reasoning was actually applied in any given clinical case. A fundamental feature of the LHS is the generation and curation of clinical evidence using electronic data sources. Such a process is critically dependent on a full transparency of how evidence is produced, maintained and consumed as a means of generating trust in the underlying system. Trust in the evidence base leads to the acceptance of responsibility for the clinical recommendations made by it which is essential if these tools are to gain widespread acceptance in the clinical community.

2.2. Data provenance

Put simply, data provenance describes what actually happened for some data entity to achieve its current form. W3C standards body defines provenance as a form of contextual resource metadata *that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility.* The Office of the National Coordinator (ONC) for Health IT describes it as *attributes about the origin of health information at the time it is first created and tracks the uses and permutations of the health information over its lifecycle*. Term *data provenance* is used to establish the focus on data entities produced in the processes.

Data provenance provides traceability by automatically capturing the trace of the research task and resulting data in a uniform and domain-independent way, thereby facilitating reproducible research. The original concept comes from the eScience and cyber-infrastructure communities, where it was used for capturing the exact parameterisations and configurations of scientific workflows that produced a particular data set [9,10]. Although the original users of provenance data were the scientific programmers creating and maintaining research workflows, the increasing number of tools and technologies available resulted in a wide array of stakeholders who can benefit from provenance information using visual front-end tools and interactive reports.

2.2.1. PROV model

The provenance technology, as defined in the W3C PROV standard [3], provides a common platform for automated capture of metadata about the data artifacts (e.g. databases, individual patient records, diagnostic recommendations), all processes that use or create those artifacts, and all actors that participate in those processes, such as clinicians, patients, researchers, or computer software. The resulting provenance data stores are typically