Web Semantics: Science, Services and Agents on the World Wide Web 29 (2014) 19–30

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

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

Domain-specific summarization of Life-Science e-experiments from provenance traces

Alban Gaignard [a,*], Johan Montagnat [a], Bernard Gibaud [b], Germain Forestier [c], Tristan Glatard [d,e]

[a] Université de Nice Sophia Antipolis / CNRS UMR7271 I3S, MODALIS team, Sophia Antipolis, France
[b] Université de Rennes 1 / INSERM U1099 LTSI, Rennes, France
[c] Université de Haute-Alsace, MIPS (EA 2332), Mulhouse, France
[d] Université de Lyon 1 / CNRS UMR5220 / INSERM U1044 / INSA-Lyon CREATIS, Lyon, France
[e] McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Canada

ARTICLE INFO

Article history:
Received 26 July 2013
Received in revised form
6 July 2014
Accepted 6 July 2014
Available online 19 July 2014

Keywords:
E-Science
Workflows
Provenance
Linked data

ABSTRACT

Translational research in Life-Science nowadays leverages e-Science platforms to analyze and produce huge amounts of data. With the unprecedented growth of Life-Science data repositories, identifying relevant data for analysis becomes increasingly difficult. The instrumentation of e-Science platforms with provenance tracking techniques provides useful information from a data analysis process design or debugging perspective. However raw provenance traces are too massive and too generic to facilitate the scientific interpretation of data. In this paper, we propose an integrated approach in which Life-Science knowledge is (i) captured through domain ontologies and linked to Life-Science data analysis tools, and (ii) propagated through rules to produced data, in order to constitute human-tractable experiment summaries. Our approach has been implemented in the *Virtual Imaging Platform* (VIP) and experimental results show the feasibility of producing few domain-specific statements which opens new data sharing and repurposing opportunities in line with Linked Data initiatives.

1. Life-Science data acquisition and production

Digital Life-Science data, ranging from molecular scale (*e.g.* proteins structural information) to human-body scale (*e.g.* radiological images) and including records as diverse as biological samples, epidemiological data, and clinical information, is acquired using many kinds of sensors. Its proper interpretation usually requires dense information on the acquisition context, the subject studied, and possibly the socio-economical environment of patients concerned. Consequently, many medical data storage and communication formats tightly associate metadata with the raw data acquired, to produce as much as possible self-contained and informative data sets. With the generalization of digital data acquisition sensors, the standardization of data acquisition formats,[1] and the online availability of Life-Science data,[2] the community has clearly turned towards the use of standard semantic data description and manipulation technologies developed in the context of the Semantic Web.[3]

To speed-up time-to-discovery in medical research, the so-called *Translational Medicine* movement reuses and relates information generated through uncoordinated multi-disciplinary data acquisition procedures and stored into very large, geographically

* Corresponding author. Tel.: +33 492965176.
E-mail addresses: alban.gaignard@cnrs.fr (A. Gaignard), johan.montagnat@cnrs.fr (J. Montagnat), bernard.gibaud@univ-rennes1.fr (B. Gibaud), germain.forestier@uha.fr (G. Forestier), tristan.glatard@mcgill.ca (T. Glatard).

http://dx.doi.org/10.1016/j.websem.2014.07.001
1570-8268/

[1] Among which the *Digital Image and COmmunication in Medicine* (DICOM—medical.nema.org) or the *Health Level Seven* (HL7—www.hl7.org) standards, just to name a few.

[2] Not only bioinformatics data is commonly available in public or research-oriented databases nowadays, but also international-scale biology and epidemiological data is published openly to boost research against health societal challenging diseases such as cancer and mental disorders.

[3] Especially through the use of taxonomies and ontologies among which the *Foundational Model of Anatomy* (FMA—http://sig.biostr.washington.edu/projects/fm) or the *Systematized Nomenclature of Medicine—Clinical Terms* (SNOMED-CT—http://www.ihtsdo.org/snomed-ct), just to name a few.

distributed data sources (*e.g.* genomic and radiological data). Annotations-aware data formats and communication standards facilitate raw data archiving at the level of each acquisition site. They pave the way toward data search, reuse and repurposing in the context of *Linked Data* [1] that underlies translational medicine, beyond the boundaries of a single discipline or community [2]. However, many different "standards" have emerged especially when linking data from different sub-disciplines. Data deluge in Life Sciences is not only a matter of volume but also a matter of diversity [3,4] as both structural heterogeneity (incompatible formats) and semantic heterogeneity (multiple terminologies and conceptualizations) are common.

To face the data deluge and facilitate resources sharing, scientists increasingly use e-Science platforms [5] dedicated to Life Sciences in order to capture raw data and transform it into well-documented data sets of interest for future exploration. Collaborative e-Science platforms are typically used to perform *in-silico* experiments, share the resources involved, and produce new valuable data (*e.g.* to evaluate a data analysis procedure onto several open databases, or to quantitatively compare several data analysis procedures through a common reference database). But to enable the reuse of (and possibly to repurpose) data in future studies, it is critical for e-Science platforms to keep track of the links between source data, produced data, and annotations associated either to the source data or the transformation process itself. This *data provenance* information facilitates data reinterpretation, data quality assessment, data processing validation, debugging, experiment reproducibility, scientific outcomes ownership control, etc. Platforms are nowadays commonly instrumented with provenance data capture.

When large data sets are manipulated, the provenance capture process generates very large annotation stores. Although provenance provides useful fine-grained and technical information on data analysis procedures, it does not ensure a better understanding of data produced from a scientist perspective due to (i) the size and the fine granularity of provenance information, (ii) the reference to technical details of the analysis pipelines, and (iii) the lack of links with relevant domain concepts. Valuable information may be available, yet deeply buried in the data stores. The first objective of this work is to *instrument data processing tools with domain-specific information* describing both the kind of data processed and the data transformation process implemented (see Section 4). Based on this captured knowledge, the second objective of this work is to analyze the dense provenance traces generated, combined with the tools and source data annotations, *to produce experiment summaries which are both human-tractable and informative for scientists* (see Section 5).

This paper proposes a methodology, leveraging Semantic Web technologies and standards, to instrument e-Science medical data processing platforms in order to capture and produce knowledge related to processed medical data. It discusses the resulting metadata deluge challenge and introduces new ways of reducing the amount of metadata generated to tractable, scientifically informative summaries through the use of domain-oriented ontologies and production rules. Concrete results are demonstrated through an implementation of this methodology in the *Virtual Imaging Platform*[4] (VIP) [6].

The remainder of this paper is organized as follows: Section 2 describes the VIP platform and exemplifies the limitations of raw provenance usage through a concrete use case. Section 3 illustrates the overall approach. Section 4 gives more details on how domain knowledge can be captured and associated to e-Science workflows and Section 5 describes how this knowledge can be used to generate experiment summaries. Section 6 provides some qualitative and quantitative experimental results. Limitations of our approach, as well as related works are discussed in Section 7 and perspectives are drawn in Section 8.

2. Platform and scenario

2.1. The VIP simulation platform

The Virtual Imaging Platform is an e-Science platform for medical image simulation. Medical image simulations combine descriptions of a medical image acquisition device (physical characteristics and parameterization), an object to image (anatomical and possibly pathological or physiological object), and a simulation scene (geometry and spatial coordinates of both the device and the object to image). The platform is multi-modal since it integrates several simulators and predefined simulation workflows for each modality (Computed Tomography, Magnetic Resonance, Positron Emission Tomography, and Ultrasound), and multi-organ since several anatomical or physiological models can be used. Simulating medical images has a variety of applications in research and industry, including fast prototyping of new devices and the evaluation of image analysis algorithms [7–9].

Performing medical image simulation is challenging for several reasons. First, simulators are complex softwares with a steep learning curve (fine parameterization, requiring a deep understanding of their physical principles) and hardly interoperable. Second, the organ models are complex, possibly involving complex anatomical/pathological characteristics, movement or longitudinal follow-up. Finally, realistic simulations are compute-intensive, and thus require dedicated computing infrastructures. VIP relies on the *European Grid Infrastructure* (EGI)[5] to support its computing and storage needs. Between October 2012 and January 2014, 6723 simulations were run, which corresponds to more than 700 CPU years, for more than 380 users originating from 40 countries.

VIP massively produces simulated data. Handling provenance in VIP is crucial to face the coherent sharing of (i) input organ models, (ii) simulator themselves, (iii) simulated data and their associated knowledge. VIP faces the issues of producing not only raw data, but also populating its simulated data repository with meaningful data. It thus needs to bridge the gap between provenance in technical simulation workflows and domain knowledge formalized with the OntoVIP domain ontology [10,11] (see Section 4.1).

2.2. Usage scenario

VIP simulators are complex and they are described as multi-step workflows to facilitate their parallelization. The enactment of medical imaging simulation workflows produces large amounts of data. Some is only intermediate data, whereas the resulting simulated data is useful for end-users. The usage scenario proposed here tracks provenance in *Sorteo* [12], one of the VIP simulation workflows, in order to address:

- *a technical concern*, allowing for workflow designers and experiment operators to more easily determine the cause of failure or abnormalities; and
- *a reliability concern*, making scientists more confident in the data produced through their experiments since the reproducibility of simulation experiments is made easier and data lineage can be controlled.

[4] http://vip.creatis.insa-lyon.fr.

[5] EGI, www.egi.eu, is a distributed multi-sciences computing platform federating hundreds of thousands of CPU cores distributed in hundreds of computing centres all over Europe and beyond.