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

Alban Gaignard, Johan Montagnat, Bernard Gibaud, Germain Forestier, Tristan Glatard

Domain-specific summarisation 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.*

Abstract

Translational research in Life-Science nowadays leverages e-Science platforms to analyse 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 provide 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.

Keywords: E-Science, Workflows, Provenance, Linked Data

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 generalisation of digital data acquisition sensors, the standardisation 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 distributed data sources (*e.g.* genomic and radiological data).

Email addresses: alban.gaignard@cnrs.fr (Alban Gaignard), johan.montagnat@cnrs.fr (Johan Montagnat), bernard.gibaud@univ-rennes1.fr (Bernard Gibaud), germain.forestier@uha.fr (Germain Forestier), tristan.glatard@mcgill.ca (Tristan Glatard)

[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.

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 analyse 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 parameterisation), 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 multimodal 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, 8, 9].

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

Performing medical image simulation is challenging for several reasons. Firstly, simulators are complex softwares with a steep learning curve (fine parameterisation, requiring a deep understanding of their physical principles) and hardly interoperable. Secondly, 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, 12] (see section 4.1).

2.2. Usage scenario

VIP simulators are complex and they are described as multi-steps 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* [11], 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.

This scenario shows that raw provenance traces can hardly be exploited by end-users since their technicality, their size and the lack of semantics hamper the interpretation of produced data from the scientist perspective.

[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.

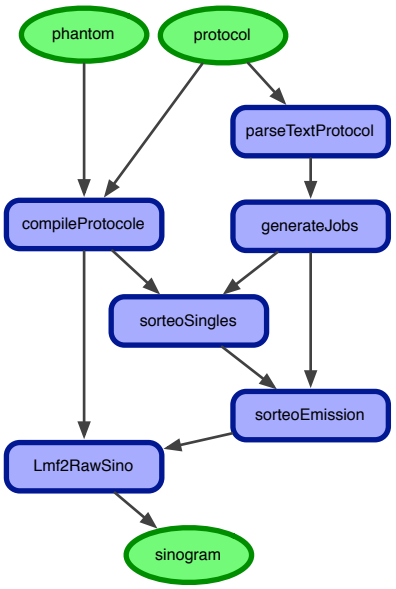

Figure 1: Graphical representation of the *Sorteo* PET medical image simulation workflow.

Sorteo is a Monte Carlo-based medical image simulator dedicated to the production of synthetic Positron Emission Tomography (PET) data. PET is a functional imaging modality, used in the field of nuclear medicine, that shows *in-vivo* quantitative metabolic activities. A simplified version of the *Sorteo* simulation workflow is presented in Figure 1. Blue boxes represent either compute-intensive activities whose executions are relocated on the EGI grid or lightweight activities executed locally. Green ellipses represent input or output data. Intermediate data sets produced by each processing step are not represented explicitly in this graph but the corresponding data flow is shown as black arrows linking computational processes.

The main workflow inputs are the *protocol*, storing all simulation parameters, and the *phantom* representing the object model to be virtually imaged. The *Sorteo* simulation workflow produces a single output, a *sinogram*, representing the simulated PET data. The core of the simulation consists in two steps:

(i) the parallel computation of "singles" through the *sorteoSingles* activity ; and
(ii) the parallel computation of the "emissions" through the *sorteoEmission* activity.

In each execution of the *Sorteo* workflow, these two activities are instantiated concurrently several times, de-

pending on the size of the simulation. The remaining activities can be considered either as pre- or post-processing steps, needed to assemble simulation parameters, or to convert data throughout the simulation workflow.

Produced data. In a single workflow execution and for a fixed set of parameters, this *Sorteo* simulator generates more than 150 data entities. Depending on workflow parameters such as the size of the virtual medical image and the number of jobs used to compute the "singles" of the Monte-Carlo simulation, a simulation workflow execution may generate a huge amount of intermediate data files (one PET "emission" file per Monte-Carlo job computing "singles"). Finally, the simulation workflow produces a single reconstructed file (the "sinogram") based on all intermediate PET "emissions". Depending on their goals, users have different interests for the data sets produced. Inspecting intermediate data such as PET "emission" may have an interest when debugging the simulation process, but these files may probably be ignored in other scenarios.

Provenance information capture. We consider in this scenario a provenance-instrumented workflow engine able to trace all fine-grained simulation activities. Provenance information is actually represented in an RDF graph and relies on the OPM ontology [26].

OPM represents causal dependencies between "things" through directed graphs. A Causal dependency is defined as a directed relationship between an *effect* (the source of the edge) and a *cause* (the destination of the edge). A node of the provenance graph might either be an *Artifact* (immutable, stateless element), a *Process* (action performed on *Artifacts* and producing new ones), or an *Agent* (entity controlling or affecting the execution of a *Process*). Graph edges represent (i) dependencies between artifacts (*wasDerivedFrom*) to track data lineage, (ii) dependencies between two processes (*wasTriggeredBy*) to track the sequence of processes, and (iii) dependencies between artifacts and processes (*used*/*wasGeneratedBy*) to track artifacts production and consumption through processes. In addition, OPM tracks the links between processes and their enactor agents through *wasControlledBy* dependencies.

As an example, Listing 1 illustrates the main provenance statements describing the execution of the last processing step of the workflow. It traces the execution of the Lmf2RawSino process. An instance of the *Process* class is created with the http://vip.cosinus.anr.fr/run-LMF2RAWSINO-1 URI, constructed from a prefix, the name of the workflow processor and a uniform unique identifier (UUID). This process execution is attached to an OPM *Account*, which represents the overall workflow execution. Note that all OPM *Artifacts* and *Processes* registered through a single workflow execution are also attached to an OPM *Account*.

```
<http://vip.cosinus.anr.fr/run−LMF2RAWSINO−1>
    a opmv:Process ;
    opmo:account <http://vip.cosinus.anr.fr/workflow−1> .
```

Listing 1: OPM statements describing the Lmf2RawSino process execution.

Listing 2 illustrates the causal "data production" dependency registered between the previous Lmf2RawSino process execution and the output sinogram. This dependency is represented by an instance of the *WasGeneratedBy* OPM class and is identified similarly to processes. This instance is linked to both the process execution through the *cause* OPM property, and the *Artifact* describing the output sinogram through the *effect* OPM property. In addition, the process input or output ports are described through the *role* OPM property linking together the data dependency and an instance of the OPM *Role* class which corresponds to the label of the process input or output port. Finally, the data production is timestamped through the OPM *time* property towards an instance of the OPM *OTime* class.

```
<http://vip.cosinus.anr.fr/wgb−1>
    a opmo:WasGeneratedBy ;
    opmo:account <http://vip.cosinus.anr.fr/workflow−1> ;
    opmo:cause <http://vip.cosinus.anr.fr/run−LMF2RAWSINO−1> ;
    opmo:effect <http://vip.cosinus.anr.fr/artifact−1> ;
    opmo:role <http://vip.cosinus.anr.fr/role−1> ;
    opmo:time <http://vip.cosinus.anr.fr/time−1> .
```

Listing 2: OPM statements describing the WasGeneratedBy dependency between the output sinogram and the Lmf2RawSino process.

Finally listing 3 describes the OPM *Artifact* corresponding to the output sinogram of the *Sorteo* PET simulation workflow. An *Artifact* instance is created. It has already been attached to the *WasGeneratedBy* causal dependency through the *effect* property of the previous listing. An *Artifact* is an abstract entity and OPM allows for associating their concrete values. The *Artifact* is thus linked to an instance of the *AValue* OPM class through the *avalue* property. Finally, a content is associated to the value through the OPM *content* property. This content finally gives the logical file name (LFN) of the sinogram, a URI locating the data on the EGI grid infrastructure. Data might be later on downloaded through a dedicated data transfer interface.

```
<http://vip.cosinus.anr.fr/artifact−1>
    a opmv:Artifact ;
    opmo:account <http://vip.cosinus.anr.fr/workflow−1> ;
    opmo:avalue <http://vip.cosinus.anr.fr/value−1> .

<http://vip.cosinus.anr.fr/value−1>
    a opmo:AValue ;
    opmo:account <http://vip.cosinus.anr.fr/workflow−1> ;
    opmo:content "lfn://lfc−biomed.in2p3.fr/grid/biomed/creatis/vip/data
     /users/rafael_silva/sorteo−2/24−01−2012_10:13:30
     /dataLMF.ccs.sino"^^<http://www.w3.org/2001/XMLSchema#anyURI> .
```

Listing 3: OPM statements describing the sinogram produced as an output of the Lmf2RawSino process.

The use of the OPM ontology leads to verbose provenance annotations. Indeed, more than 14 RDF statements (timestamping has not been represented) are necessary to represent a single data item production in the *Sorteo* workflow. This is mainly due to the reification of all dependencies, leading to complex paths between provenance entities (we consider here only a single data production).

Finally, OPM statements illustrated above represent technical information such as the location of produced files in a distributed computing infrastructure, the name of the processing tools involved in simulation experiments, or time-stamping. They represent precise and fine-grained information, beneficial when inspecting logs of medical imaging simulations, however, they do not convey any domain-specific information such as simulation modality or high-level parameters, useful for medical imaging experts.

Needs for concise and domain-specific provenance. Although precise provenance statements are definitely necessary for technical workflow refinement or debugging, the size, the fine granularity of provenance and its lack of links with domain-specific concepts makes it unmanageable from a scientist perspective, possibly running workflows on large input datasets. As an example, a single run of the *Sorteo* workflow leads to a large OPM technical provenance graph composed by 4523 nodes and 15154 edges.

To address this issue, we propose to distinguish between two levels of provenance information. First, fine-grained domain-agnostic provenance (represented through standards provenance models such as OPM), useful for technical workflow refinement or debugging. Second, coarse-grained domain-specific provenance, representing concise domain-specific statements resulting from production rules relying on the VIP medical image simulation ontology. These produced statements will finally constitute "semantic experiment summaries" in which a minimal set of statements link together simulation experiment results to experiment parameters through the OntoVIP [10, 12] domain ontology.

3. Global approach

Our approach is based (i) on knowledge capture, *i.e.* data and services semantically annotated with a domain ontology (see Section 4), and (ii) on knowledge production, by applying production rules to annotate the processed data with new concise domain-specific statements finally assembled into semantic experiment summaries (see Section 5).

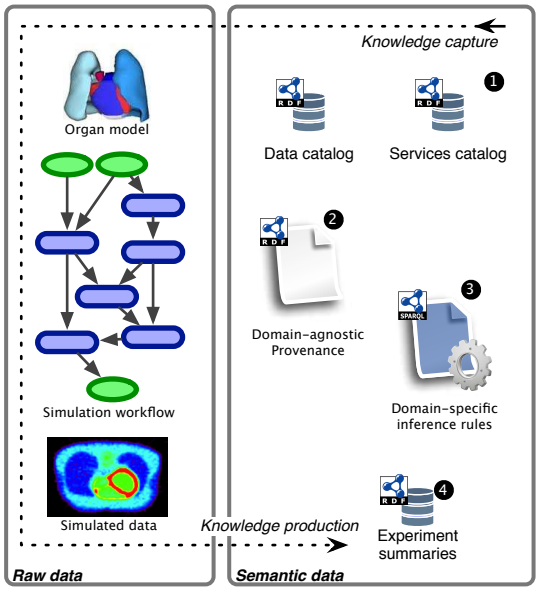

Figure 2: Knowledge capture and production to produce semantic experiments summaries from medical imaging workflow executions.

Figure 2 illustrates the proposed approach. The left part of the figure focuses on *raw data*: organ model and medical image simulator selection, simulator parameterisation and execution, and simulated data production.

The right part of the figure focuses on semantic data, and illustrates our approach to produce semantic experiment summaries.

- First we rely on (i) semantically annotated input data (organ models), and (ii) semantically annotated services actually composed into simulation workflows ❶.
- Then, domain-agnostic provenance ❷ is tracked on the fly at workflow runtime and represented

through a standard model (OPM in the current implementation) .

- When the workflow successfully produced a simulated data, the set of available domain-specific production rules ❸ are applied. Each applicable rule involves semantic service annotations, and produces new domain-specific statements.
- These new statements finally constitute the semantic experiment summaries and populate a dedicated catalog ❹.

The joint querying of the catalogs for organ models, simulators, and experiment summaries, published in the platform following Linked Data principles, opens interesting perspectives in terms of simulated data and organ models sharing and reuse.

4. Knowledge capture in e-Science workflows

To propagate knowledge from domain ontologies to data produced through e-Science workflow executions, concepts defined through a domain ontology must be associated to data processing services syntactical elements. If we consider the last processing step of the *Sorteo* medical image simulation workflow (Figure 1) for example, it consumes two input parameters that share the same syntactic type. The first parameter is typed with a URI representing the imaging protocol file path, the second parameter is also typed with a URI which represents the path of the directory containing all generated emissions to be effectively reconstructed by "Lmf2RawSino". These syntactic types do not precisely characterize input data. To have a clear understanding of the transformation realized on input and the output data, both the service itself and its parameter should refer to concepts of a domain ontology.

We rely on the OntoVIP [10, 12] ontology (section 4.1) to model the medical image simulation domain and the OWL-S [19] generic service ontology (section 4.2) to semantically annotate services composed into workflows. We also highlight the issue of ambiguous semantic service annotations and propose in this knowledge capture process, to pay a particular attention in distinguishing *Role* and *Natural* concepts when annotating service parameters (section 4.3).

4.1. Overview of the OntoVIP medical image simulation ontology

OntoVIP was developed to facilitate the sharing and automated processing of information managed within the VIP Virtual Imaging Platform. OntoVIP provides a coherent conceptual framework, grounded on the DOLCE (Descriptive Ontology for Language and Cognitive Engineering) foundational ontology [13]. The ontology was built through intensive interviews with researchers involved in image simulation. It took almost a year to reach a consensual modeling and several incremental versions were produced. OntoVIP is publicly available through the BioPortal[6].

It includes medical information object models (called for short *'Object models'* in the following) whose sharing and reuse are essential since such models are hard to build from scratch, and can often be easily derived from existing ones. This part of the ontology involves a taxonomy of object models, highlighting their content: *e.g.* geometrical phantom object model or biological object model, whether they contain some external agent (*Object model with external agent*) or some foreign body (*Object model with foreign body*), their compatibility with simulators (*i.e.* whether they specify the physical properties of objects required with a particular class of simulator, *e.g.* CT[7] simulation compatible model), whether they are static (*Static object model*) or dynamic, *i.e.* modeling a moving object (*Moving object model*) or an object undergoing some evolution in time (*Longitudinal follow up object model*). This taxonomy is complemented by entities describing the content of the object models' geometry files (*e.g. 3D voxel matrices* or *meshes*) to relate them to classes of real-world objects (*e.g. Anatomical object*, *Pathological objects*, *Foreign body objects*). The latter classes were extracted from existing ontologies such as FMA [14] (Foundational Model of Anatomy), RadLex [15] (Radiology Lexicon), MPATH [16] (Mouse pathology).

OntoVIP also includes a detailed taxonomy of simulated data, *i.e.* data resulting of the execution of some medical image simulation software. This taxonomy involves three major semantic axes. The first is related to imaging modality (*e.g.* CT simulated data, MR[8] simulated data); the second makes a distinction between non-reconstructed data and reconstructed data (*i.e.* images); the former are further categorized into classes denoting the spatial or spatiotemporal organization of the data (*e.g.*, list-mode data, sinogram, set of signals, set of projection images); and finally the third distinguished between static simulated data and dynamic simulated data.

Simulated data are the result of the execution of some medical image simulator, *i.e.* software whose func-

[6] OntoVIP: http://bioportal.bioontology.org/ontologies/3253
[7] X-ray Computed Tomography.
[8] Magnetic Resonance.

tion is to perform medical image simulation. Medical image simulators and medical image simulations are further categorized depending on imaging modalities (*i.e.* CT, MR, US[9], PET[10]). Medical image simulators are composed of simulator components addressing the different stages of a simulation: pre-processing (implemented by pre-processing simulator component), core simulation (implemented by core simulation simulator component) and post-processing (implemented by post-processing simulator component). OntoVIP models the relationships between simulated data and object models, and between simulated data and parameter sets or parameters; such relationships (derivedFromModel, derivedFromParameterSet, derivedFromParameter, respectively) are of key importance with regards to the domain-specific modeling of data lineage.

OntoVIP was developed based on OntoNeuroLOG, an ontology developed during the NeuroLOG project[11] for supporting the sharing of heterogeneous and distributed medical images and image processing tools in neuroimaging [17, 18]. The OntoVIP ontology is used in the VIP software to support the annotation and querying of models, as well as the annotation and querying of simulated data and of the data processing actions that actually produced this data.

4.2. Semantic service annotation

To complete the *Knowledge Capture* task on medical image simulation workflows, Semantic Services associate concepts of the OntoVIP ontology to the service descriptors composing simulation workflows. The field of Semantic Web Services aims at exploiting semantic web technologies to enhance service oriented architectures and thus e-Science workflow environments. Through a rich, formal and standard semantic description, benefits are expected both at workflow design-time, when discovering, composing and mediating services, and at workflow run-time, when linking back processed data to semantic service annotations. Our approach focuses on the latter to produce human-tractable and informative enough semantic experiment summaries.

Several initiatives have been targeting the standardization of semantic service description such as, for extended frameworks, OWL-S [19], WSMO [20], FLOWS [21], or for lighter approaches, SAWSDL [22] or WSMO-Lite [23]. Although SAWSDL has been proposed by the W3C as a recommendation in 2007, no consensus clearly emerged, and OWL-S and SAWSDL provide good compromises for semantically annotating e-Science workflow components.

We reused the *OWL-S Profile* ontology concepts to describe semantically the key processing steps of medical image simulation workflows, in terms of functionality, input and output parameters. These descriptions have been bridged to the OntoVIP domain ontology through *refers-to* properties. As an example, the "Lmf2RawSino" *refers to* the *image-reconstruction-simulator-component* OntoVIP class to describe its functionality, and *refers to* the *PET-Sinogram* class to describe the produced data through its output parameter.

4.3. Role concepts in semantic service descriptions

We showed in [24] that only considering the intrinsic *Nature* of parameters would possibly lead to ambiguous semantic service annotations.

When exploiting a workflow execution in terms of provenance information, it may not be possible to identify a unique path in the data production chain, due to some parameters of a particular processing step, sharing the same *Nature*. For example, two input parameters may share the same *Nature* (*e.g.* Magnetic Resonance modality) but be considered differently from a data processing perspective. The first input parameter may be considered as a reference data, and the second one may be considered as the data to be analyzed or transformed. More generally, data can play different roles in the context of a data processing tool.

Without this contextual knowledge specifying how data are related, through one or more parameters, to a specific data processing step, it is difficult to deduce domain-specific information from the workflow executions and their associated provenance. Few approaches such as FLOWS *fluents* [21] or BioCatalogue *functions* [25], may be used to make the distinction between the nature of service parameters and their role from the service perspective. We also pay a particular attention in making the distinction between *Role* and *Nature* concepts at domain ontology design time. *Roles* can then be used, to disambiguate semantic service descriptions finally enabling reasoning and the production of new domain-specific statements from workflow executions.

5. Producing semantic experiment summaries from e-Science workflow runs

Based on disambiguated semantic services and provenance traces, reusable production rules instrumenting domain ontologies enable the production of

[9] Ultrasound.
[10] Positron emission tomography.
[11] NeuroLOG project: http://neurolog.unice.fr

new domain statements. Due to compute-intensive tasks, a single workflow execution may lead to a huge amount of fine-grained provenance information, as explained in Section 2.

Section 5.1 first introduces the OPM-O domain-agnostic provenance ontology. Section 5.2 then proposes to use domain-specific production rules (through SPARQL CONSTRUCT queries) to propagate domain knowledge, from raw provenance traces, to the processed data through new domain-specific statements, finally assembled into semantic experiment summaries.

5.1. Domain-agnostic provenance ontologies

The *Open Provenance Model* [26] initiative (OPM) is a community effort aiming at homogenizing the expression of provenance information on the wealth of data produced by e-Science applications. OPM broadly addresses workflow environment interoperability through a standardized representation and easier exchanges of provenance information. It also eases the development of tools to process such provenance information, and finally facilitates the reproducibility of e-Science experiments.

Provenance ontologies are crucial initiatives helping in precisely tracking provenance information, which open interesting interoperability and reproducibility perspectives, in the context of e-Science applications. However, these standardization initiatives do not consider any specific domain. When presenting such provenance information to e-Scientist, we face two main issues :

- *Granularity:* e-Scientists are often not aware of all the constituents of a particular workflow and they generally consider workflows as grey-boxes in which only part of the produced data is of interest. Systematic provenance tracking and representation through domain-agnostic provenance ontologies leads to large fine-grained bunch of information, hampering the interpretation of workflow results.

- *Abstraction:* e-Scientists are nowadays used to attach precise meaning (through domain ontology) to their data or processing tools, to enhance their representation and sharing. However, standard provenance ontologies are domain-agnostic. Domain-agnostic provenance ontologies are not sufficient to properly interpret and share processed data. This requires in addition, leveraging a domain-oriented ontology.

To tackle these issues, we propose to design domain-specific production rules. They address (i) the automated semantic annotation of generated raw data, and (ii) the semantic summarisation of e-Science experiments through few domain-specific statements.

5.2. Reusable and service independent rules to produce new concise domain-specific statements

New domain-specific statements are produced from e-Science workflow executions using (i) domain ontologies, (ii) domain-agnostic provenance information, and (iii) domain-specific production rules. SPARQL is the standard language dedicated to Semantic Web graphs querying. Although, SELECT is its most popular query form, for graph data selection, CONSTRUCT queries allow for producing new RDF statements when a graph pattern is matched. It can thus be considered as a production rule composed with an *antecedent* (an *"If"* condition), its WHERE clause, and a *consequent* (a *"Then"* consequence), the CONSTRUCT clause.

As a detailed example, we propose the production rule illustrated in Listing 4. Its WHERE clause identifies a sub-graph into the full OPM-O provenance statements while its CONSTRUCT clause produces new domain-specific statements describing, in a concise form, the whole simulation experiment. More precisely, this CONSTRUCT query augments the initial RDF graph with new triples leveraging the OntoVIP ontology. They state (i) the nature (*type*) and the location (*is-stored-in-file*) of the input parameters and output data, (ii) the nature and relations of the produced medical images with respect to the input organ model and the simulation workflow (*derives-from-model, is-a-result-of-at*, etc.), and (iii) the nature and global parameters of the simulation workflow (*uses-as-model-in-simulation*, etc.). Being concise and domain-specific, these statements semantically annotate the produced raw data, and helps e-scientists interpret data produced along their experimental campaigns and link the simulation parameters and components to the produced data.

- *Lines 13 to 27* represent the new statements resulting from the application of this rule. They only involve domain-specific entities (*vip-model*, *vip-simulation* and *vip-simulated-data* prefixes of the OntoVIP ontology) whereas the WHERE clause of the query only involves OPM-O entities. These new statements semantically describe the nature of input parameters (*medical-image-simulation-object-model*, *simulation-parameter-set*) and simulated output medical images (*PET-sinogram*). In

```
PREFIX rdf: <http://www.w3.org/1999/02/22−rdf−syntax−ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf−schema#>
PREFIX opmo: <http://openprovenance.org/model/opmo#>
PREFIX opmv: <http://purl.org/net/opmv/ns#>
PREFIX ws: <http://www.irisa.fr/visages/team/farooq/ontologies/web−service−owl−lite.owl#>
PREFIX iec: <http://www.irisa.fr/visages/team/farooq/ontologies/iec−owl−lite.owl#>

PREFIX vip−model: <http://vip.cosinus.anr.fr/vip−model.owl#>
PREFIX vip−simulation: <http://vip.cosinus.anr.fr/vip−simulation.owl#>
PREFIX vip−simulated−data: <http://vip.cosinus.anr.fr/vip−simulated−data.owl#>

CONSTRUCT {
        ?inPhantom rdf:type vip−model:medical−image−simulation−object−model
        ?inPhantom vip−model:is−stored−in−file ?cInPhantom

        ?inProtocole rdf:type vip−simulation:simulation−parameter−set
        ?inProtocole vip−model:is−stored−in−file ?cInProtocole

        ?out vip−model:derives−from−model ?inPhantom
        ?out vip−simulation:derives−from−parameter−set ?inProtocole
        ?out rdf:type vip−simulated−data:PET−sinogram
        ?out vip−model:is−stored−in−file ?cOut
        ?out vip−simulation:is−a−result−of−at ?wf

        ?wf rdf:type vip−simulation:PET−simulation
        ?wf vip−simulation:uses−as−model−in−simulation ?inPhantom
        ?wf vip−simulation:uses−as−parameter−in−simulation ?inProtocole
} WHERE {
        ?agent (iec:refers−to/rdf:type)
            <http://vip.cosinus.anr.vip.fr/vip−simulation.owl#image−reconstruction−simulator−component> .
        ?wcb opmo:cause ?agent .
        ?wcb opmo:effect ?x .
        ?x rdf:type opmv:Process .
        ?wgb opmo:cause ?x .
        ?wgb opmo:effect ?out .

        ?agent2 (iec:refers−to/rdf:type)
            <http://vip.cosinus.anr.vip.fr/vip−simulation.owl#parameters−generation−simulator−component> .
        ?wcb2 opmo:cause ?agent2 .
        ?wcb2 opmo:effect ?y .
        ?y rdf:type opmv:Process .

        ?used1 opmo:cause ?inPhantom .
        ?used1 opmo:effect ?y .

        ?used2 opmo:cause ?inProtocole .
        ?used2 opmo:effect ?y .

        ?used1 opmo:role/rdfs:label ?techRolePhantom .
        ?used2 opmo:role/rdfs:label ?techRoleProtocole .

        ?agent2 ws:has−input ?inPortPhantom .
        ?inPortPhantom (iec:refers−to/rdf:type)
            <http://vip.cosinus.anr.fr/vip−model.owl#geometrical−phantom−object−model> .
        ?inPortPhantom rdfs:comment ?techRolePhantom .

        ?agent2 ws:has−input ?inPortProtocole .
        ?inPortProtocole (iec:refers−to/rdf:type)
            <http://vip.cosinus.anr.fr/vip−model.owl#quality−procedure−dataset> .
        ?inPortProtocole rdfs:comment ?techRoleProtocole .

        ?x opmo:account ?wf .

        ?inPhantom opmo:avalue ?vInPhantom .
        ?vInPhantom opmo:content ?cInPhantom .

        ?inProtocole opmo:avalue ?vInProtocole .
        ?vInProtocole opmo:content ?cInProtocole .

        ?out opmo:avalue ?vOut .
        ?vOut opmo:content ?cOut .
}
```

Listing 4: Production rule based on a SPARQL CONSTRUCT query to associate the input phantom to the produced output sinogram resulting from an execution of the *Sorteo* simulation workflow.

addition, these new annotations represent the nature of the simulation (*i.e.* Positron Emission Tomography) through the use of the *PET-simulation* class, and the relation between the medical object imaged and the produced simulated image (*derives-from-model* property). They are particularly useful because they involve medical imaging concepts and relations forming part of the OntoVIP ontology.

- *Lines 29 to 33* identify a process execution, its corresponding service description through an *?agent* instance, and the achieved class of action through the *iec:refers-to* property. In this rule, the class of action is an image reconstruction.

- *Lines 34 to 35* identify the output *Artifact* (*?out*) through an instance of the *WasGeneratedBy* causal dependency (*?wgb*). This dependency is linked to the *Process* (*?x*) through an *opmo:cause* property, and to the output *Artifact* through an *opmo:effect* property. The value and content are associated to the *Artifact* through the *opmo:avalue* and the *opmo:content* properties (lines 70 to 71).

- *Lines 37 to 41* identify a process execution realizing a parameters generation action, similarly to lines 29 to 33.

- *Lines 43 to 60* identify the *Artifacts* used as input of a process execution realizing a parameters generation action. Additionally, the *?role* characterizing how the *Artifact* has been used by the process is identified. It identifies the parameters in the semantic service description associated to the process (line 49 and 50).

- *Lines 52 to 60* finally join the service description of (*?agent2*) to the process execution (*?y*) through the label associated to the input port (*?role*), this input port referring to a geometrical phantom (lines 53 to 55).

Due to the fine granularity of the OPM-O provenance ontology, the graph pattern to be matched is large. Developing this kind of production rules is time-consuming and error-prone, it is thus important to foster the reusability of such rules.

By relying on domain specific taxonomies to describe services we enhance the rule reusability. As an example, if we consider a new version of the *Sorteo* workflow where the last process has been updated to *Lmf2RawSino_v2*, the same production rule can be reused. Indeed, we consider that its implementation is completely different (technical parameters may have changed) but its functionality is still the same. Since the semantic description of *Lmf2RawSino_v2* is subsumed by the semantic description of *Lmf2RawSino*, and the production rule involves semantic description of *Lmf2RawSino*, the same production rule can successfully be applied to also annotate the results of *Lmf2RawSino_v2*.

Through the use of (i) fine-grained technical provenance and (ii) domain-specific production rules, we presented a method exploiting domain ontologies, not only at workflow design-time, but also at workflow runtime. Our method propagates domain knowledge on processed data to finally constitute semantic experiment summaries. In the following section, we propose experimental results showing the interest of generating few domain-specific statements, to enhance workflow results interpretation, especially in the context of Linked Data.

6. Results

6.1. Experimental setup

The VIP simulation platform hosts a semantic catalog of organ models which associates the set of raw source files with the set of semantic annotations describing each model. It leverages the OntoVIP medical image simulation ontology and enables advanced querying on available models. VIP consumes organ models through the MOTEUR data-intensive workflow manager [27] coupled to the European Grid distributed computing Infrastructure (EGI[12]) to produce simulated data. It keeps records of all running and completed simulations, enabling simulated data search, post-analysis and reuse. More details on the VIP platform are available in [6].

Through the work presented in this paper, simulated data entities are annotated with OntoVIP-based semantic information linking them to (i) an input organ model, (ii) an input parameter sets, and (iii) a brief description of the overall simulation workflow. To achieve this result, the VIP platform was instrumented with:

- A semantic catalog of composable simulator components;

- A fine-grained OPM provenance traces generation plugin for the MOTEUR workflow manager to capture on-the-fly provenance information;

[12]EGI: http://egi.eu

- Domain-specific production rules (SPARQL CONSTRUCT) aiming at automating the semantic annotation and the summarisation of medical image simulation workflows for 4 modalities (Computed Tomography (CT), Magnetic Resonance (MR), Ultrasound (US) and Positron Emission Tomography (PET)); and

- A graphical viewer for the resulting experiment summaries providing a tabular representation of the simulated data catalog and their relations with input organ models and medical image simulators.

Finally, both the catalog of simulation services (describing the function of data processing steps and the nature and the role of their parameters) and the catalog of organ models (describing for example anatomical knowledge) are used with domain-agnostic provenance traces to populate the simulated data catalog with new experiment summary statements as illustrated in Figure 2.

Technically, the semantic repository and reasoner are built upon open source libraries such as Apache JENA for RDF data persistency, Corese/KGRAM[13] for Semantic Web querying and reasoning, Apache Commons and Log4J for helper classes, and JSPF for a simple java plugin framework.

Two experiments are proposed below to assess the scalability of our approach when producing semantic experiment summaries, and to show the usability of these summaries, especially in the context of Linked Open Data.

6.2. *Semantic experiment summaries for scalable e-Science data annotation*

New statements resulting from production rules provide high-level and concise "semantic experiment summaries". We consider experiment summaries as high-level descriptions since they only involve domain-specific classes and properties defined in the OntoVIP ontology, compared to the generic and technical entities provided by the OPM provenance ontology. We also consider them concise since for *Sorteo*, only 12 statements might be produced, compared to thousands of OPM statements produced through our provenance-instrumented workflow engine.

Figure 3 illustrates the experiment summary resulting from the execution of the *Sorteo* medical imaging workflow. Green ellipses represent input or output data, the blue ellipse represents the *Sorteo* workflow shown as a "black box", and red rectangles represent VIP ontology classes. The production rule presented in Listing 4 automates the semantic annotation of the output sinogram and the corresponding input phantom and input protocol. Dashed green arrows represent the new inferred statements. For instance, the output sinogram is related to its corresponding input phantom through the *vip-model:derives-from-model* property (Listing 4, line 19). The nature of the sinogram is also determined through the *is-a* property towards the VIP class *PET-Sinogram* (Listing 4, line 21).

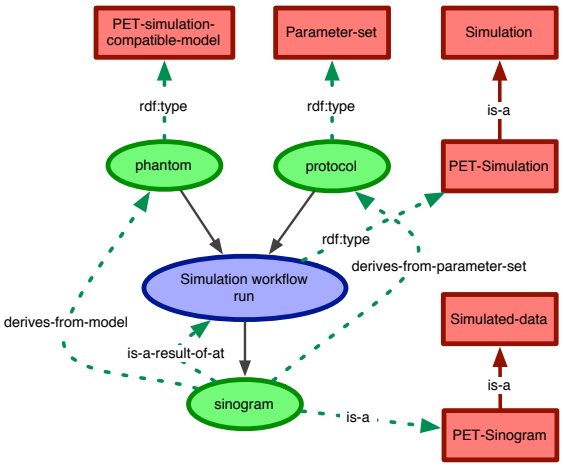

Figure 3: New produced statements (dashed arrows) constituting the semantic experiment summary.

Scalability. Since technical fine-grained OPM provenance information is useful at workflow design-time and workflow debug-time, it is temporarily stored in a short-term semantic repository. To produce new domain-specific statements, only the provenance information related to a single execution, and the service descriptions are needed. We finally store in a long-term repository the few "experiment summary" statements.

When running the *Sorteo* medical imaging workflow, only 12 RDF triples are produced as experiment summary when more than 1400 OPM-O RDF triples are recorded through the provenance-instrumented workflow engine. Between December 2012 and June 2013, 136 medical image simulations have been summarised. These summaries consists in 3114 domain-specific RDF triples. They represent only 3.5% of the size of the corresponding full OPM-O provenance graphs (87587 triples). As an illustration, Figure 4 gives a visual idea on the content of the VIP long-term repository storing the experiment summaries. It shows that the VIP platform has been mainly used, during this period, for CT,

[13] Corese/KGRAM: http://wimmics.inria.fr/corese

MR and US simulation (lot of instances for the *CT-simulation*, *MR-simulation*, and *US-simulation* classes). This kind of graphical representation also shows that a single organ model ("organes.pegs4dat") has been significantly used as input model for CT simulations. To extract more precise information with respect to the VIP platform usage, it is still possible to perform quantitative SPARQL queries on the RDF experiment summaries. As an example, SPARQL *count* queries involving OntoVIP domain-specific entities show that the VIP long-term repository is composed of 3114 summary triples, and represents 39 US simulations, 31 CT simulations, 64 MR simulations, and 3 PET simulations. Another quantitative query shows that the "organes.pegs4dat" organ model has been used in all the 31 CT simulations, which represents 22% of the overall simulations. It shows that this organ model has been intensively used in the VIP platform to perform CT simulations.

Performance. In terms of performance, both the capture of raw provenance and the production of experiment summaries is negligible compared to the processing time of raw data on a dedicated computing infrastructure. Over 233 simulation workflow runs, we measured a mean summary production time of 2 seconds with a standard deviation of 1.1 seconds. The mean ratio of summary production time over workflow execution time is 0.76%.

We pay special attention to scalable data production through (i) the materialization of few produced statements into a long-term simulated data catalog, and (ii) short-term fine-grained OPM provenance datasets stored independently and available for workflow design and debug concerns.

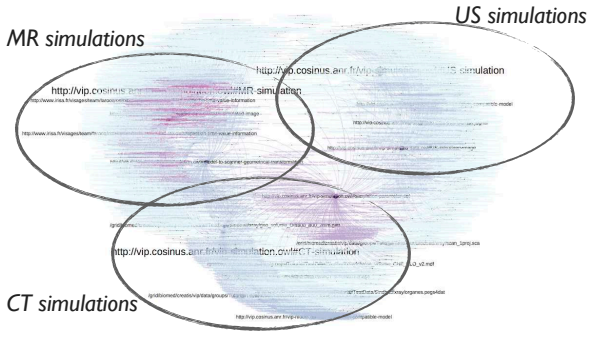

Figure 4: Visual content of the long-term simulated data catalog showing three main groups of medical image simulations.

6.3. Semantic experiment summaries exploitation

6.3.1. Linked Data querying

Based on provenance information, our approach automates the annotation of e-Science workflow results with domain-specific concepts, finally assembled, following Link Data [1] principles, into semantic experiment summaries. These summaries can be combined with external data sources such as FMA [14], the Foundational Model of Anatomy. We exemplify in Listing 5 a SPARQL query leveraging three kinds of interlinked data, VIP simulated data, VIP organ models, and FMA anatomical concepts. This query aims at retrieving simulated data from VIP organ models which contain brain white matter, as defined in the FMA ontology.

```
PREFIX mo: <http://vip.cosinus.anr.fr/vip-model.owl#>
PREFIX partic: <http://www.irisa.fr/visages/team/farooq/ontologies/
      particular-owl-lite.owl#>
PREFIX iec: <http://www.irisa.fr/visages/team/farooq/ontologies/iec-owl-
      lite.owl#>
PREFIX fma: <http://sig.uw.edu/fma#>

SELECT ?dataFile ?dataClass ?anat WHERE {
  ?organModel rdf:type mo:medical-image-simulation-object-model .
  ?organModel partic:has-for-proper-part-during ?x .
  ?x rdf:type mo:anatomical-object-layer .
  ?x partic:has-for-proper-part-during ?y .
  ?y rdf:type mo:anatomical-object-layer-part .
  ?y iec:refers-to ?anat .

  ?anat rdf:type fma:White_matter_of_neuraxis .

  ?dataArtifact mo:derives-from-model ?model .
  ?dataArtifact mo:is-stored-in-file ?dataFile .
  ?dataArtifact rdf:type ?dataClass .
FILTER (CONTAINS(?organModel,?model))
}
```

Listing 5: SPARQL query exploiting both the VIP organ model catalog and the newly populated simulated data catalog and FMA concepts

The first six triple patterns appearing in the WHERE clause aim at searching for VIP organ models (*medical-image-simulation-object-model*), and their anatomical constituents (*anatomical-object-layer* and *anatomical-object-layer-part*). Reference anatomical concepts are retrieved through the *refers-to* property and the *?anat* variable. Then, the next triple pattern specifies the FMA anatomical concept to be matched: brain white matter (*White_mater_of_neuraxis*). Finally, the last three triple patterns aim at retrieving, from the simulated data catalog, data files (*is-stored-in-file*) and their associated medical image modality simulated from organ models (*derives-from-model*) including white matter.

6.3.2. Simulated data catalog

One of the main objectives of the VIP platform is to ease the setup of medical image simulation experiments.

The VIP web-based graphical user interface hides the complexity of the underlying simulation workflows and the distributed data management.

In this context, the proposed semantic experiment summaries have been directly exploited, through SPARQL queries, to populate a catalog of simulated data. This simulated data catalog finally helps e-Scientists in searching or retrieving simulated medical images, based on their modality, on the organ models used in the simulation, or on the simulation parameters. In this catalog, simulated data are linked to the simulations used to produce them, so that users can retrieve the exact parameters and logs on request.

7. Discussion

7.1. Related works

With regards to generic provenance ontology standardization, OPM recently made a step further. It evolved through a W3C standardization process towards the PROV-* specifications[14]. PROV-O [34] is an OWL specification of the W3C provenance data model (PROV-DM). Evolving from basic OPM-O provenance representation to PROV-O is almost direct. There is a mapping between the root classes: *Artifact* ↔ *Entity*, *Process* ↔ *Activity*, *Used* ↔ *Usage*, or *WasGeneratedBy* ↔ *Generation*. A noticeable enhancement is the definition of simple properties for *usage* and *generation* causal dependencies. Whereas these dependencies must be reified with OPM-O, leading for instance to two triples which link a process instance to an artifact through an intermediate instance of the *opmo:WasGeneratedBy* class[15], only a single triple is needed with PROV-O (the instantiation of the dependency is not required anymore).

In addition, PROV-O extends OPM-O with some classes and properties especially useful in the context of e-Science workflows. For instance, PROV-O introduces the notion of *Plan* to describe the context of execution of an *Activity*, which can be seen as a set of instructions, as a recipe, or a workflow. Another interesting extension is the *alternateOf* property aiming at representing several aspects of the same thing. For instance, in medical imaging, it would be well adapted to link several datasets resulting from data conversion tools.

[14] http://www.w3.org/TR/prov-primer

[15] see the example of Listing 2

Madougou and coworkers propose in [35] a provenance-based approach aimed at analyzing the e-BioInfra e-Science platform usage and identifying the causes of application failures. From a *post mortem* analysis of the MOTEUR workflow enactor logs, the proposed system populates a relational SQL backend with OPM provenance statements, queried through HQL (Hibernate Query Language). The natural graph representation of provenance is buried into a relational representation, and the system cannot benefit from graph-based querying languages such as SPARQL. The system addresses the technical characterization of workflow executions through a statistical analysis of fine-grained domain-agnostic provenance.

We rather focus on result interpretation from the e-scientist perspective by leveraging domain-specific ontologies and preserving the underlying graph structure of provenance, thus allowing for graph-based querying and reasoning through Semantic Web technologies.

The Wings/Pegasus environment [36] addresses through semantic reasoning on application-level constraints, the generation of valid and execution-independant workflows, to be enacted over distributed computing infrastructures. The system proposed is able to produce both application-level and execution provenance. Wings/Pegasus uses a proper OWL ontology to model application-level provenance data and uses a provenance tracking catalog, based on a relational database, to record execution provenance. Two languages are thus required to query provenance data, SPARQL for design-time application-level (and thus domain-specific) provenance, and SQL for run-time domain-agnostic execution provenance.

Rather than using two representations for execution-level and application-level provenance, we rely on RDF, for a graph-based representation of these two levels. Wings/Pegasus also attaches domain knowledge to workflow templates which is an interesting perspective to reduce the design complexity of our production rules.

In *Janus*, [37] introduce semantic provenance as technical provenance graphs coupled with domain knowledge. The main objective is to enhance the usefulness of provenance graphs in responding to typical user queries. Semantic provenance was first introduced by [38]. Missier and coworkers propose with *Janus* a domain-aware provenance model by extending the Provenir upper-level ontology [39] grounded to BFO [41] (Basic Formal Ontology) concepts, and a prototype implementation within the Taverna work-

flow workbench. The modeling of domain entities relies on four ontologies registered in NCBO, the National Center for Biomedical Ontologies, namely the BioPAX (dedicated to the modeling of biological pathways), the National Cancer Institute (NCI) Thesaurus, the Foundational Model of Anatomy (FMA) and the Sequence ontology. Once web services composed into Taverna workflows are semantically annotated, simple inference rules for each service execution are responsible for the propagation step-by-step of semantic annotations to the produced domain-agnostic provenance, thus providing new domain-specific provenance. To answer provenance queries, a specific transitive closure implementation was proposed based on low-cost SPARQL Ask queries.

Janus is definitely the closest approach to our proposal for generating e-Science experiment summaries. The main differences are the use of OPM-O and the medical imaging ontologies OntoVIP grounded to DOLCE in our work, compared to Provenir and biomedical ontologies in *Janus*. To address scalability issues, we propose to make a clear distinction between short-term fine-grained domain-agnostic provenance and produced long-term domain-specific provenance through semantic experiment summaries. *Janus* extends domain-agnostic provenance with domain specific statements, which requires to manage in a single dataset the large amount of fine-grained provenance.

Also addressing the exploitation of e-Science workflows from an end-user perspective, Alper and coworkers analyze in [40] why raw provenance traces are difficult to exploit and share in the context of data publication. They motivate the distillation of raw provenance into more usable and focused provenance, hiding the noise of less significant processing steps or intermediate data. They propose an interesting solution based on knowledge capture which consists in annotating at design-time, workflow templates or "Motifs". They address a similar objective which consists in generating "origin-annotations" on input parameters and propagating them, at run-time, onto produced data through table representations. In addition, they propose to create workflow summaries based on "Motifs" annotations, however, the bindings between produced and annotated data with "origin-annotations" and workflow summaries is not obvious.

Our approach addresses similar objectives and is in line with the analysis of Alper and coworkers. We try to provide an integrated way of producing semantic experiment summaries involving coarse-grained domain-specific annotations which interlink produced/analyzed data to (i) input parameters, and (ii) design-time annotations of processing services. We rely on Semantic Web standards to ease the publication of experiment summaries through Linked Data principles.

7.2. Added value and limitations

Semantic web services. Services involved in e-Science workflows are generally described through detailed WSDL descriptors, possibly allowing for syntactic validation. However, RESTful services have recently been largely adopted due to lighter deployment and better flexibility. As an example, the KEGG[16] WSDL services were decommissioned in december 2012 and migrated to REST interfaces. No consensus emerged to semantically annotate RESTful services, but SA-REST [28], which relies on RDFa to describe a service with RDF triples embedded into a companion HTML document, or [29], bridging WADL descriptors to OWL-S appear as potential solutions, both in line with our approach.

Production rule design. Although the design of production rules can be complex, the simulation workflows deployed in production in the VIP platform keep stable. Rules are therefore reused all along the platform life-time. More precisely, it took 2 persons/month to design the 18 production rules, grouped into 4 modalities : 1 rule for Ultrasound, 1 rule for PET, 1 rule for CT, and 15 rules for MR (with very slight variations due to similar workflow structures). As an example, during 6 months of VIP operation, we recorded 137 medical imaging simulations in which 39 of them were Ultrasound. The single US rule has been reused 39 times for this modality. Similarly, during the same period, the rule summarizing CT simulations has been reused 31 times.

When developing production rules, the order of triple patterns may have a significant impact on performance. Their design is thus crucial and they should be reused as much as possible when workflows evolve. This is made possible by the loose coupling between production rules and services descriptions. The proposed rules adapt to several service implementations as long as they are semantically annotated with the same domain ontology concepts (or sub-concepts). However, they remain highly dependent on the structure of scientific workflows. Workflow evolutions would require adapting the production rules. Abstract (or conceptual/template) workflow initiatives such as the conceptual workflows introduced in [30, 31] could help in the design of

[16] Kyoto Encyclopedia of Genes and Genomes

production rules. Indeed, fine-grained workflow structures could be hidden by higher level conceptual workflow elements and production rules could be attached to these abstract workflow components instead of being attached to fine-grained provenance statements, thus enhancing their reuse.

More practically, the MOTEUR workflow designer could be extended to generate, based on the workflow structure and selected elements, the summarization rules. In terms of production rules correctness, this extension could validate the rules proposed by the workflow designer through a set of SPARQL queries that would check some domain constraints. As an example, a validation query would check that each produced data has a domain-specific type, and is linked to input data through specific properties (*derives-from-model* and *derives-from-parameter-set*).

Graph summarisation techniques have also been proposed to reduce graph complexity and to extract informative content [32]. The genericity of these approaches is appealing and would reduce the design cost of domain-specific rules. However, it remains to be seen to what extent the graph structure criterions used in summarisation are relevant in the context of e-Science workflows.

Usability and quality. Our approach aims at enhancing the usability of data produced through e-Science workflows, and more precisely, medical imaging workflows involved in the VIP platform. Both usability and quality are considered. Workflow designers can exploit raw fine-grained OPM-O provenance information while designing and debugging workflows. But due to provenance traces size and genericity, it is not aimed at being directly exploited by scientists. Through the proposed semantic experiment summaries, we aim at enhancing the confidence of scientists in the quality of their experiments by providing concise domain-specific annotations describing the produced data and coarse-grained relations between the data produced and the experiment parameters.

A user-oriented evaluation would be necessary to validate our approach and study the possible usage of experiment summaries. It would also bring valuable inputs on how e-scientists search for their simulated data, and if the proposed approach foster sharing of simulation data/models. Currently, these summaries are used to populate the simulated data catalog exposed to end-users through the VIP web portal. Platform logs show that for the last 6 months (December 2013 - April 2014), 137 experiment summaries were produced and the simulated data catalog has been viewed 68 times.

Sharing of experiment summaries. To tackle the interpretation of possibly massive data production in the context of e-Science workflows, we automate the generation of semantic experiment summaries. We produce these summaries from OPM-O provenance datasets. These experiment summaries represent new concise domain-specific statements in the sense that we associate the produced data to concepts and relation of the OntoVIP domain ontology. These summaries make sense for e-scientists if they are aware of the OntoVIP ontology and the medical image simulation domain. To enhance the sharing of experiments summaries outside this community, we could also rely on extensions of the PROV-O provenance ontology to represent these summaries.

However, although it is possible to extend the PROV-O ontology with domain-specific taxonomies, these extensions may raise ontology design issues, typically if the domain ontology (*e.g.* OntoVIP) is grounded to foundational ontologies such as DOLCE or BFO [41] (Basic Formal Ontology). Grounding PROV-O to a foundational ontology would allow smart articulations with domain ontologies also grounded to foundational ontologies such as BIOTOP [42] (Top-Domain ontology for the life sciences) or OBI [43] (Ontology of Biomedical Investigation). Garijo and coworkers proposed P-PLAN [33] an extension of PROV-O to represent workflows and also stressed the interest of grounding it to DOLCE. However, the counterpart would certainly be a consequent design effort needed to bridge together PROV-O with foundational ontologies.

Simulated data reuse perspective. The SPARQL query illustrated in Listing 5 shows the relevance of producing domain-specific provenance information in e-Science platforms by joining interlinked data catalogs. Not only does it allow accurate search for simulated data but it also enhances the sharing, reuse and repurposing of existing simulated data. Reusing already computed simulated data could save a lot of computing and storage resources, and opens interesting perspectives towards smarter simulation platforms (less CPU, memory, and time). Re-exploitating medical image simulation experiments from the perspective of anatomical models also opens interesting educational perspectives (*e.g.* learning medical imaging through simulation, quick understanding of the parameters impact on simulated data).

8. Conclusion & future works

E-Science experimental platforms use data-intensive workflows to massively process data. Tracking work-

flow provenance is crucial to improve reproducibility of e-experiments and confidence in both data and processing chains. Due to its size, its fine-granularity and the lack of relations with domain ontologies, the exploitation of raw provenance traces is however humanly intractable.

Our approach enables domain-specific knowledge capture and generation in the context of medical image simulation workflows. It promotes a clear delineation between *Role* and *Natural* concepts in domain ontologies to disambiguate the semantic annotation of service parameters, thus providing more accurate semantic service descriptions. It proposes a way of augmenting domain ontologies with inference rules that produce human-tractable and informative experiment summaries out of fine-grain provenance trace sets.

Results show that it is possible to instrument the main medical imaging workflows of the VIP platform with domain-specific provenance summarisation rules to produce few domain-specific statements. Besides, representing and querying experiment summaries through Semantic Web technologies opens exciting sharing and repurposing perspectives, especially in the context of Linked Open Data.

We consider two main continuations for this work. First, to link domain-specific experiment summaries with the fine-grained raw traces used for their generation, so that detailed technical execution traces can be retrieved when necessary. Second, to improve the genericity of our approach. The methodology proposed in this paper could easily be applied to other disciplines massively producing data (*e.g.* Bioinformatics) but it may require a modeling effort to instrument domain ontologies with proper production rules. We plan to study how generic graph summarisation techniques or abstract graph representations could help in producing experiment summaries at a lower design cost.

Acknowledgments

This work is funded by the French National Research Agency (ANR) under grant ANR-09-COSI-03 and the CNRS interdisciplinary mission MASTODONS under program CrEDIBLE. We thank the European Grid Initiative and "France-Grilles".

We would also like to thank Olivier Corby and Catherine Faron Zucker for their support and advises regarding the Corese/KGRAM Semantic Web engine.

References

[1] C. Bizer, T. Heath, T. Berners-Lee, Linked Data - the story so far, Int. J. Semantic Web Inf. Syst. 5 (3) (2009) 1–22.

[2] P. Fox, J. A. Hendler, Semantic eScience: encoding meaning in next-generation digitally enhanced science., in: T. Hey, S. Tansley, K. M. Tolle (Eds.), The Fourth Paradigm, Microsoft Research, 2009, pp. 147–152.

[3] C. Goble, R. Stevens, State of the nation in data integration for bioinformatics, J. of Biomedical Informatics 41 (5) (2008) 687–693. doi:10.1016/j.jbi.2008.01.008.

[4] C. A. Goble, D. D. Roure, The impact of workflow tools on data-centric research., in: T. Hey, S. Tansley, K. M. Tolle (Eds.), The Fourth Paradigm, Microsoft Research, 2009, pp. 137–145.

[5] T. Hey, A. Trefethen, Cyberinfrastructure for e-Science, Science 308 (5723) (2005) 817–821. doi:10.1126/science.1110410.

[6] T. Glatard, C. Lartizien, B. Gibaud, R. Ferreira Da Silva, G. Forestier, F. Cervenansky, M. Alessandrini, H. Benoit-Cattin, O. Bernard, S. Camarasu-Pop, N. Cerezo, P. Clarysse, A. Gaignard, P. Hugonnard, H. Liebgott, S. Marache, A. Marion, J. Montagnat, J. Tabary, D. Friboulet, A Virtual Imaging Platform for multi-modality medical image simulation, IEEE Transactions on Medical Imaging (TMI) 32 (1) (2013) 110– 118.

[7] A. Gilliam, S. Acton, Echocardiographic simulation for validation of automated segmentation methods, in: Image Processing, 2007. ICIP 2007. IEEE International Conference on, Vol. 5, 2007, pp. V – 529–V – 532. doi:10.1109/ICIP.2007.4379882.

[8] M. Prastawa, E. Bullitt, G. Gerig, Simulation of brain tumors in MR images for evaluation of segmentation efficacy. Medical Image Analysis 13 (2) (2009) 297–311.

[9] G. Wagenknecht, H.-J. Kaiser, T. Obladen, O. Sabri, U. Buell, Simulation of 3D MRI brain images for quantitative evaluation of image segmentation algorithms (2000) 1074–1085doi:10.1117/12.387612.

[10] B. Gibaud, G. Forestier, H. Benoit-Cattin, F. Cervenansky, P. Clarysse, D. Friboulet, A. Gaignard, P. Hugonnard, C. Lartizien, H. Liebgott, J. Montagnat, J. Tabary, T. Glatard, Ontovip: An ontology for the annotation of object models used for medical image simulation, in: Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on, 2012, p. 110. doi:10.1109/HISB.2012.35.

[11] A. McLennan, A. Reilhac, M. Brady, SORTEO: Monte carlo-based simulator with list-mode capabilities, in: Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE, 2009, pp. 3751 –3754. doi:10.1109/IEMBS.2009.5334536.

[12] G. Forestier, A. Marion, H. Benoit-Cattin, P. Clarysse, D. Friboulet, T. Glatard, P. Hugonnard, C. Lartizien, H. Liebgott, J. Tabary, B. Gibaud, Sharing object models for multi-modality medical image simulation: A semantic approach, in: Computer-Based Medical Systems (CBMS), 2011 24th International Symposium on, 2011, pp. 1 –6. doi:10.1109/CBMS.2011.5999167.

[13] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, WonderWeb Deliverable D18. The WonderWeb Library of Foundational Ontologies and the DOLCE ontology (December 2003).

[14] C. Rosse, J. L. V. Mejino Jr, The Foundational Model of Anatomy ontology, in: A. Burger, D. Davidson, R. Baldock (Eds.), Anatomy Ontologies for Bioinformatics, Vol. 6 of Computational Biology, Springer London, 2008, pp. 59–117. doi:10.1007/978-1-84628-885-2_4.

[15] S. Kundu, M. Itkin, D. A. Gervais, V. N. Krishnamurthy, M. J.

Wallace, J. F. Cardella, D. L. Rubin, C. P. Langlotz, The IR Radlex Project: an interventional radiology lexicon–a collaborative project of the Radiological Society of North America and the Society of Interventional Radiology., J Vasc Interv Radiol 20 (7 Suppl) (2009) S275–7.

[16] P. Schofield, J. Sundberg, B. Sundberg, C. McKerlie, G. V. Gkoutos, The mouse pathology ontology, MPATH; structure and applications, Journal of Biomedical Semantics 4 (1) (2013) 1–8. doi:10.1186/2041-1480-4-18.

[17] L. Temal, M. Dojat, G. Kassel, B. Gibaud, Towards an ontology for sharing medical images and regions of interest in neuroimaging, J. of Biomedical Informatics 41 (5) (2008) 766–778.

[18] B. Gibaud, G. Kassel, M. Dojat, B. Batrancourt, F. Michel, A. Gaignard, J. Montagnat, NeuroLOG: sharing neuroimaging data using an ontology-based federated approach. AMIA Symposium (2011) 472–80.

[19] D. Martin, M. Burstein, D. Mcdermott, S. Mcilraith, M. Paolucci, K. Sycara, D. L. Mcguinness, E. Sirin, N. Srinivasan, Bringing semantics to web services with OWL-S, World Wide Web 10 (3) (2007) 243–277. doi:10.1007/s11280-007-0033-x.

[20] D. Roman, J. de Bruijn, A. Mocan, H. Lausen, J. Domingue, C. Bussler, D. Fensel, WWW: WSMO, WSML, and WSMX in a nutshell, 2006, pp. 516–522. doi:10.1007/11836025_49.

[21] M. Gruninger, R. Hull, S. McIlraith, A short overview of flows: A First-order Logic Ontology of Web Services, IEEE Data Engineering Bulletin. 31 (3) (2008) 3–7.

[22] J. Farrell, H. Lausen. Semantic Annotations for WSDL and XML Schema [http://www.w3.org/tr/sawsdl] [online] (August 2007).

[23] T. Vitvar, J. Kopecky, J. Viskova, D. Fensel, WSMO-Lite Annotations for Web Services, in: 5th European Semantic Web Conference (ESWC2008), 2008, pp. 674–689.

[24] A. Gaignard, J. Montagnat, B. Wali, B. Gibaud, Characterizing semantic service parameters with Role concepts to infer domain-specific knowledge at runtime, in: International Conference on Knowledge Engineering and Ontology Development, KEOD 2011, Paris, France, 2011.

[25] J. Bhagat, F. Tanoh, E. Nzuobontane, T. Laurent, J. Orlowski, M. Roos, K. Wolstencroft, S. Aleksejevs, R. Stevens, S. Pettifer, R. Lopez, C. A. Goble, BioCatalogue: a universal catalogue of web services for the life sciences, Nucleic Acids Research 38 (suppl 2) (2010) W689–W694. doi:10.1093/nar/gkq394.

[26] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan, J. V. den Bussche, The Open Provenance Model core specification (v1.1), Future Generation Computer Systems 27 (6) (2011) 743–756.

[27] T. Glatard, J. Montagnat, D. Lingrand, X. Pennec, Flexible and efficient workflow deployment of data-intensive applications on grids with MOTEUR, International Journal of High Performance Computing Applications (IJHPCA) Special Issue on Workflows Systems in Grid Environments 22 (3) (2008) 347–360.

[28] A. P. Sheth, K. Gomadam, J. Lathem, SA-REST: Semantically interoperable and easier-to-use services and mashups. IEEE Internet Computing 11 (6) (2007) 91–94.

[29] O. F. F. Filho, M. A. G. V. Ferreira, Semantic Web Services: A RESTful Approach, in: IADIS International Conference WWWInternet 2009, IADIS, 2009, pp. 169–180.

[30] N. Cerezo, J. Montagnat, Scientific Workflow Reuse through Conceptual Workflows in: 6th Workshop on Workflows in Support of Large-Scale Science(WORKS'11), ACM, Seattle, WA, USA, 2011.

[31] D. Garijo, O. Corcho, Y. Gil, Detecting common scientific workflow fragments using templates and execution provenance, in: Seventh ACM International Conference on Conference on Knowledge Capture, Banff, Canada, 2013.

[32] X. Zhang, G. Cheng, Y. Qu, Ontology summarization based on RDF sentence graph, in: Proceedings of the 16th international conference on World Wide Web, WWW '07, ACM, New York, NY, USA, 2007, pp. 707–716. doi:10.1145/1242572.1242668.

[33] D. Garijo, Y. Gil, Augmenting PROV with Plans in P-PLAN: Scientific Processes as Linked Data, 2012.

[34] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. PROV-O: The PROV Ontology, W3C recommendation, http://www.w3.org/TR/prov-o, april 2013.

[35] S. Madougou, S. Shahand, M. Santcroos, B. van Schaik, A. Benabdelkader, A. van Kampen, S. Olabarriaga, Characterizing workflow-based activity on a production e-infrastructure using provenance data, Future Generation Computer Systems 29 (8) (2013) 1931 – 1942.

[36] J. Kim, E. Deelman, Y. Gil, G. Mehta, V. Ratnakar, Provenance Trails in the Wings/Pegasus System, Concurrency and Computation: Practice and Experience 20 (5) (2008) 587–597.

[37] P. Missier, S. Sahoo, J. Zhao, C. Goble, A. Sheth, Janus: from Workflows to Semantic Provenance and Linked Open Data, in: IPAW-10, 2010.

[38] S. S. Sahoo, A. Sheth, C. Henson, Semantic Provenance for eScience: Managing the Deluge of Scientific Data, IEEE Internet Computing 12 (4) (2008) 46–54. doi:10.1109/MIC.2008.86.

[39] S. S. Sahoo, A. Sheth, Provenir ontology: Towards a Framework for eScience Provenance Management, in: Microsoft eScience Workshop, 2009.

[40] P. Alper, K. Belhajjame, C. A. Goble, P. Karagoz, Enhancing and abstracting scientific workflow provenance for data publishing, in: Proceedings of the Joint EDBT/ICDT 2013 Workshops, ACM, New York, NY, USA, 2013, pp. 313–318. doi:10.1145/2457317.2457370.

[41] P. Grenon and B. Smith. SNAP and SPAN: Towards dynamic spatial ontology. *Spatial Cognition and Computation*, 4(1):69–103, 2004.

[42] E. Beisswanger, S. Schulz, H. Stenzhorn, and U. Hahn. BioTop: An upper domain ontology for the life sciences: A description of its current structure, contents and interfaces to OBO ontologies. *Appl. Ontol.*, 3(4):205–212, 2008.

[43] R. Brinkman, M. Courtot, D. Derom, J. Fostel, Y. He, P. Lord, J. Malone, H. Parkinson, B. Peters, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, L. Soldatova, C. Stoeckert, J. Turner, J. Zheng, and the OBI consortium. Modeling biomedical experimental processes with OBI. *Journal of Biomedical Semantics*, 1(Suppl 1):S7, 2010.