CHARACTERIZING SEMANTIC SERVICE PARAMETERS WITH ROLE CONCEPTS TO INFER DOMAIN-SPECIFIC KNOWLEDGE AT RUNTIME

Alban Gaignard, Johan Montagnat
CNRS / UNS, I3S lab, MODALIS team, Sophia-Antipolis, France

Bacem Wali, Bernard Gibaud
INSERM / INRIA / CNRS / Univ. Rennes 1, IRISA Unit VISAGES U746, Rennes, France

Keywords: Semantic web services, Role modeling, Reusable inference rules, Scientific workflows.

Abstract: E-Science platforms leverage Service Oriented Architecture (SOA) principles to deliver large catalogs of data processing services and experiments description workflows. In spite of their growing success, the usability of these platforms is hampered by their catalogs size and the domain-specific knowledge needed to manipulate the services provided. Relying on domain ontologies and semantic services to enhance the understanding and usability of e-Science platforms, our contribution is twofold. First, we propose to delineate role concepts from natural concepts at domain ontology design time which leads to a neuroimaging role taxonomy, making explicit how neuroimaging datasets are related to the data analysis services. Then we propose to exploit, at workflow runtime, provenance information extended with these domain roles, to infer new meaningful semantic annotations. Platform semantic repositories are thus transparently populated, with newly inferred annotations, through the execution of e-Science workflows. A concrete example in the area of neurosciences illustrates the use of role concepts to create reusable inference rules.

1 INTRODUCTION

Semantically representing information has become a *de facto* technique to enrich e-Science experimental platforms with domain-specific knowledge. This approach aims at facilitating platforms usage, sharing of experimental data and results, and experiments themselves, to finally foster collaborations among large user groups. Conceptualizing domain knowledge, ontologies became a cornerstone for the underlying Information Systems, as they are built upon controlled vocabularies, logical constraints and inference rules.

Generally relying on Service Oriented Architectures (SOA), e-Science experimental platforms provide tools dedicated to the publication, the identification, and the invocation of data processing services. However the technical description of services (*e.g.* using WSDL) does not provide any understanding on the nature of the information processed nor on the operations applied. Exploiting catalogues of data processing services, *e.g.* to design flows of services (workflows) requires a clear understanding of how data is processed and the nature of the data transformation implemented by the services. Today, users are expected to have acquired this knowledge, which limits the platforms usability to a restricted number of experts.

In this context, and relying on ontologies, semantic (web) services tend to explicit the understanding of (i) the nature of processed data and (ii) the nature of the information processing applied to benefit, both at experiment design-time and runtime, from the knowledge on the services manipulated. Different levels of semantic information can be distinguished:

1. Generic information, related to the technical description of services (*e.g.* semantic service descriptions based on OWL-S) or related to the service invocation which can later be used to produce provenance traces (*e.g.* following the Open Provenance Model).

Gaignard A., Montagnat J., Wali B. and Gibaud B..
CHARACTERIZING SEMANTIC SERVICE PARAMETERS WITH ROLE CONCEPTS TO INFER DOMAIN-SPECIFIC KNOWLEDGE AT RUNTIME.
DOI: 10.5220/0003648200590070
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2011), pages 59-70
ISBN: 978-989-8425-80-5

2. Domain-specific information related to the nature of the information processing realized by a service invocation and the nature of the data manipulated (e.g. taxonomies describing the nature of *Dataset* and *Dataset-processing*). This knowledge can be used to validate service invocations, by ensuring that the expected types of data are used when invoking a service.
3. Domain-specific information related to the *Role* played by the data involved in the service execution, from the service point of view. This knowledge is needed both to ensure coherency of service invocations, and to reason on the service invocation effect on the data processed.

Leveraging existing ontologies to describe generic information as well as domain-specific nature of data and processing tools, this paper focuses on the third level of semantic information. The proposed approach tackles 3 aspects of semantic services manipulated:

- It clarifies the bindings between service descriptions and domain concepts through a taxonomy of domain-specific *Roles*.
- It enables the coherency of service workflows design.
- It makes it possible to infer new knowledge along platform exploitation. This last point is achieved by describing reusable domain-specific knowledge inference rules associated to specific natures of processing. The application of such rules on a semantic database containing traces of services invocation enriches the experimental platforms with new valuable expert information.

We rely on the NeuroLOG platform (Montagnat et al., 2008) to implement the concepts and support experiments reported in this work. NeuroLOG is a distributed environment designed to support the setup of multi-centric studies in neurosciences. The OntoNeuroLOG ontology (Temal et al., 2008) was designed in the context of the platform development to enhance the sharing of neuroimaging data and associated data analysis services.

The remainder of this paper is organized as follows. Section 2 presents some background on role modeling and how it could be related to existing initiatives in the Semantic Web Services area. Section 3 motivates our approach through a small neuroimaging workflow example. Ontologies supporting our work are briefly presented in section 4. The benefits of relying on *Role concepts* when designing a domain ontology are exposed in section 5, followed with section 6 briefly illustrating how we complemented our workflow environment with semantic web technologies. We finally discuss and conclude our approach in section 7.

2 BACKGROUND INFORMATION

2.1 Role Modeling

In conceptual modeling, it is now agreed to separate several categories of concepts, for instance those characterizing the nature of an entity from those characterizing their relations to each others. Henriksson *et al.* propose a methodology based on the design of role-based ontologies, extending standard ontologies, to enhance ontology modularization and reusability. They promote a clear delineation between *Natural Types* and *Role Types* (Henriksson et al., 2008) : *"In role modeling, concepts that can stand on their own are called natural types, while dependent concepts are called role types"*. Sowa (Sowa, 1984) first introduced *Natural Types* to describe what is essential to the identity of an individual, and *Role Types* to describe temporal or accidental relations to other individuals. The methodology proposed by Henriksson *et al.* consists in (i) identifying the natural types of the domain, (ii) identifying accidental or temporary relationships between individuals and ensuring that role models are self-contained (for reusability) and finally (iii) defining bridge axioms to bind role types to natural types (or to link individuals through properties defined in the role model). This approach is particularly interesting in our context since in Life Science ontologies, the design effort generally focuses on the first step. Moreover, e-Science experimental platforms are generally data-driven and well supported by ontologies describing the nature of data. But few efforts concentrate in making explicit the knowledge relating data to their analysis services more deeply than just using information on data nature.

2.2 Web Service Ontologies

Semantically enhanced e-Science experimental platforms usually rely on standard generic service ontologies to describe data analysis services. The following paragraphs briefly describe major service ontologies and how they consider relations between data and services.

The *OWL-S Profile* ontology (Martin et al., 2007), one of the three ontologies forming the OWL-S proposal, aims at describing what the annotated service does. The service *Profile* presents a high-level interface of the service through the properties *hasParam-*

eter, hasInput, hasOuput. These properties link the *OWL-S Profile* ontology to the *OWL-S Process* ontology (aimed at describing the service internal behavior) which defines the service parameters (*Parameter* class) and their subclasses (*Input* and *Output* classes). The type of parameters is given through the *parameterType* property of the *Process* ontology and specifies the classes/datatypes the value of the parameter belongs to. According to the OWL-S specification, nothing is said regarding how these parameter values are related to the service process and as a consequence, these types should be considered as *natural types* as they are defined by Sowa (Sowa, 1984). To specify the relationship of parameter values to the process, it should be beneficial to rely, through the *parameterType* property, on a role ontology designed according to the methodology proposed by Henriksson *et al.*.

FLOWS (Gruninger et al., 2008; Battle et al., 2005) specifies a first-order logic ontology for Web Services. It aims at enabling reasoning on the semantics of services and their interactions. FLOWS has largely been influenced by OWL-S but in addition, it addresses interactions with business process industry standards such as BPEL. FLOWS differs from OWL-S by properly handling messages as core concepts. Messages are defined in FLOWS by a *message_type*, characterizing the type of the content, and a *payload*, the content itself. FLOWS defines also three relations to relate atomic process invocations to messages they consume as input or they produce as output: *produces*, *reads*, and *destroy_message*. The relations are very generic and do not characterize more precisely the consumption/production of messages through domain-specific entities. However, FLOWS proposes the *described_by* relation to associate a *fluent* to a message. *Fluents* are used to model "changing" parts of the world. The *described_by* relation aims at providing information on how the content of the message impact the service invocation while consuming/producing it. Intuitively, since *role types* are defined by Sowa as accidental (or evolving during time) relationships between entities, FLOWS's *fluents* could be a way to model how data are interpreted by analysis services through *Roles*.

WSMO (Roman et al., 2006) is based on *Orchestration* to describe the internal behavior of services, and on *Choreography* to describe their external behaviors. The *Choreography* of a service is described through the importation of a domain ontology, which defines the choreography state signature. This signature specifies, among other things, the service inputs and outputs as instances of the imported ontology. WSMO is a rich service modeling and enacting framework but it does not cover precisely the characterization of how processed or produced data are related to services in terms of roles. Relying on external ontologies, WSMO service interface remains compatible with any domain ontology designed using a clear separation between *natural types* and *role types*.

SAWSDL (Kopecký et al., 2007) is the W3C recommendation to semantically annotate WSDL and XML Schema documents specifying standard Web Services. These documents are bound to semantic entities through the *modelReference* XML attribute. The value assigned to a *modelReference* comprises a set of zero or more URIs identifying concepts in an ontology. Again, this specification does not bring anything new to separate the *natural type* of the annotated WSDL message from how it is related to the Web Service (its *role type*). However, depending on the availability of an ontology of roles, *modelReference* attributes could be used to bind *role types* to service parameters.

Originating from the WSMO initiative, WSMO-Lite (Vitvar et al., 2008) is built upon SAWSDL and is a lightweight bottom-up approach, to semantically describe Web Services and to enable reasoning on (i) their associated semantic annotations, and (ii) their interactions. Since WSMO-Lite uses SAWSDL to bridge domain-specific ontologies with the service description, roles types should be considered as an external feature, coming from the design of the domain ontology.

2.3 Semantic Workflow Environments

Being based on either standard service ontologies, or home-made approaches, the following paragraphs describe initiatives aiming at enhancing service discovery, in the context of workflow environments.

The METEOR-S (Sheth et al., 2008) research project is a major initiative in the Semantic Web Services area. The approach is based on a peer-to-peer middleware to address service discovery and publication. SAWSDL is used for both services annotation (through *modelReferences*) and data mediation (through schema *lifting/lowering*).

Built upon the ^{my}Grid ontology (Wolstencroft et al., 2007), a bioinformatics service and domain ontology, FETA (Lord et al., 2005) is a service discovery framework characterized by a light-weight semantic support and a semi-automatic approach. Three main actors are distinguished in this framework: both *knowledge engineers* and *service annotators* provide semantic enhanced web services consumed by *scientists*. Also built upon the ^{my}Grid ontology, the BioCatalogue (Bhagat et al., 2010) initiative is a community-driven, and curated service registry aiming at guiding users into a jungle of web services through the registration and annotation of web services and the browsing of resulting annotated web services. Several kinds of annotations are available going from free text, to tags or ontology terms. BioCatalogue allows, among other kind of annotations to *operationally* (e.g. infrastructure, runtime constraints) or *functionally* describe a service. Functional annotation covers information related to what the service does, but also its function and the format of input or output data. The function annotation of data with regard to a given web service seems to be close to *Role types* previously introduced but few information is available to precisely describe this kind of annotation.

The BioMOBY project aims at providing interoperability for biological data centers and analysis centers. SAWSDL has been used in this context and this real-world application is one of the few existing initiatives (Gordon and Sensen, 2008). The focus is on interoperability and therefore on schema mapping annotations of SAWSDL, implemented through XSLT stylesheets. The entry-point is a SAWSDL Proxy servlet, in front of a web service provider, a semantic registry, and a schema mapping server. As a continuation of this initiative, the SADI project (Withers et al., 2010) proposes guidelines and best-practices to enhance semantic service discovery at workflow design time. Semantic services are indexed in the catalog through the new set of RDF properties describing the resulting new semantic features associated to input data. The service discovery is based on searches over data properties consumed as input and over the produced new properties. This approach also aims at reducing ambiguity of search queries through more precise properties, describing the relationships between input and output data. We propose to address such relationships at domain ontology design time, through a taxonomy of *Roles*, clearly identifying the role of data with regard to their analysis services.

3 MOTIVATING USE CASE

3.1 Image Registration Workflow

The workflow illustrated in Figure 1 represents a typical image registration process commonly encountered in neurosciences workflows. It consists in superimposing two medical images acquired independently into the same coordinate system. The sample registration process is decomposed into two steps. First, the registration itself consists in calculating, from the input brain MRI and a brain atlas, a geometrical transformation expressed by a transformation matrix. Second, the resampling step effectively aligns the input brain MRI by applying the transformation expressed through the registration matrix.

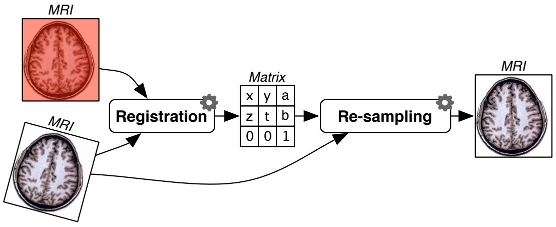

Figure 1: A typical neuroimaging workflow mixing several nature of data and processing.

In spite of its apparent simplicity, this workflow is interesting for the following reasons. First, this workflow mixes two services of different nature, whose meaning has been agreed upon within the image processing community. In other words, the knowledge about what kind of underlying treatment is clear for the community and is generally not explicit at the tooling level. Second, this workflow consumes and produces data of several natures (medical images, transformation matrix) expressed through raw files at the tooling level. Again, these files have a precise meaning from the user community point of view, with regard to their content. Finally, the first step of the workflow takes two files as input, sharing the same nature, both are brain MRIs, but they play different roles from the processing tool perspective. The first one is used as the reference image for the registration process (atlas) whereas the second one is used as the floating (moving) image. This knowledge is hidden at the tooling level, and even for domain experts, the variability of tools makes their configuration not trivial.

3.2 Enriching the Semantic Repository with Valuable Annotations

Relying on a semantic data repository together with a reasoning engine, we consider in this paper a methodology for producing and deducing new meaningful facts from the user community perspective. For example, considering the result of the registration workflow presented in Figure 1, it should be interesting to retrieve, the atlas used in the registration. More generally, our approach tends towards the propagation of the effect of a service (or a sub-part of the workflow) to the produced data. For instance, we would like to automate the generation of a fact saying that "this dataset can be superimposed with this dataset", because some processing tools might require that their inputs have beforehand been registered.

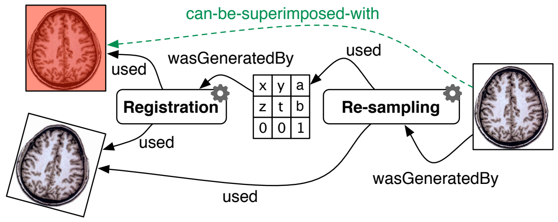

Figure 2: Linking data and processes through generic and domain-specific relations.

Figure 2 illustrates the semantic relations established between entities involved in the sample workflow, *i.e.* data and services. Black arrows are relations that can be created on-the-fly during each service invocation. It states data production and consumption knowledge. Beyond linking together data and processes (*i.e.* capturing provenance information), we want to rely on both an ontology and a reasoning engine to infer relevant domain information using rules expressing domain knowledge. For instance the dashed arrow represents such information derived using a domain-specific rule embedding domain knowledge about the overall registration process.

As it will be shown in Section 5, this kind of knowledge inference is possible if the services semantic description is rich enough to properly define the *Roles* of processed data in the context of services invocation. In addition, such high-level semantic description can be used to validate the coherency of flows of services. Before entering into the details, the next Section describes the ontologies on which this work is grounded.

4 SUPPORTING ONTOLOGIES

4.1 Domain Ontology

The OntoNeuroLOG ontology was developed to provide common semantics for information sharing throughout the NeuroLOG system. Indeed, the ultimate goal of NeuroLOG was to allow the successful sharing of neuroimaging resources provided by collaborating actors in the field of neuroimaging research, the term resources covering both neuroimaging data (such as images) as well as image processing programs, shared as services. This ontology is used as a reference to query and retrieve heterogeneous data, thanks to the mediation system, as well as to annotate consistently the shared services, *i.e.* denote what sort of processing such services actually achieve and what data they accept as input and produce as result.

OntoNeuroLOG was designed as a multi-layer application ontology, relying on a number of core ontologies modeling entities that are common to several domains. The whole ontology relies on DOLCE (Descriptive Ontology for Language and Cognitive Engineering), a foundational ontology that provides both the basic entities (at the top of the entities' taxonomy) and a common philosophical framework underlying the whole conceptualization. The ontology was designed according to the OntoSpec methodology (Kassel, 2005), which focuses on the writing of semi-formal documents capturing rich semantics. This is followed by an implementation of a subset of the ontology in OWL, the web ontology language. The definition of this subset and the choice of the relevant OWL dialect take into account the specific needs of the application. Two subsets of OntoNeuroLOG were used in the context of this work, the ontology of *Dataset* and the ontology of *Dataset processing*, introduced hereafter.

4.1.1 Dataset Sub-ontology

Datasets are *Propositions* (i.e. *Non physical endurants*) that represent the content of data files used in neuroimaging. The taxonomy of *Datasets* is organized according to several semantic axes. The first denotes what facet of the subject is explored, *e.g.* *Anatomical datasets* explore the subject's anatomy whereas *Metabolic datasets* explore brain metabolic processes. The second axis classifies *Datasets* according to some imaging modality, such as *Computed Tomography* (CT), *Magnetic resonance* (MR), *Positron emission tomography* (PET). This axis includes the numerous sub-modalities met, *e.g.*, in MR imaging such as *T1-weighted MR dataset*, *Diffusion-weighted MR dataset*, etc. The third axis focuses on Datasets that result from some kind of post-processing, such as *Reconstructed datasets*, *Registration datasets*, *Segmentation datasets*, etc.

Datasets may bear properties of *Representational objects* (since *Propositions* are *Representational objects*), such as *'refers to'*, which denotes the ability to refer to any kind of *Particular*. This property can be used to refer, *e.g.* to the *Subject* (*i.e.* the patient) concerned by a particular *Dataset*. For instance, a property called *'can be superimposed with'* was introduced to relate two *Datasets* that can be superimposed with each other, such as a *Segmentation dataset* (*i.e.* an object mask obtained through a segmentation procedure) and the original dataset from which it was obtained.

4.1.2 Dataset-processing Sub-ontology

Dataset processings are Conceptual actions (i.e. *Perdurants*) that affect *Datasets*. The taxonomy of Dataset processings covers the major classes of image processing met in neuroimaging, such as: *restoration*, *segmentation*, *filtering*, *registration*, *re-sampling*, etc. Axioms attached to each *Dataset processing* class usually denote which classes of *Datasets* are being processed or result of the corresponding processing. For example, a *Reconstruction 'has for data'* some *Non-reconstructed dataset* and *'has for result'* some *Reconstructed dataset*; a *Segmentation 'has for result'* some *Segmentation dataset*.

4.2 Ontology of Web Services

In addition, an ontology was defined to describe Web Services grounded to the DOLCE foundational concepts. It introduces the notions that are classically involved in WS specifications such as the notions of interface (*ws-interface*), operation (*ws-operation*), service inputs and outputs (*input/output-variable*). Besides, the model introduces a *'refers to'*property to establish relationships with the classes of data processing that a particular *ws-operation* implements (such as *rigid-registration* or *segmentation*), as well as with the classes (natural types) of entity that the input and output variable actually represent.

4.3 OPM Ontology

The *Open Provenance Model* (Moreau et al., 2011) initiative (OPM) aims at homogenizing the expression of provenance information on the wealth of data produced by e-Science applications. Among other things, OPM enables the exchange of provenance information between several workflow environments. It eases the development of tools to process such provenance information, and finally facilitates the reproducibility of e-Science experiments.

OPM is materialized through a natural language specification and three formal specifications: an XML schema (OPMX), an OWL ontology (OPMO) and a controlled vocabulary, with simpler OWL constructs (OPMV). OPM defines directed graphs representing causal dependencies between "things". 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). The nodes of the provenance graph might be either an *Artifact* (immutable, stateless element), or a *Process* (actions performed on an *Artifact* and producing new ones), or an *Agent* (entity controlling or affecting the execution of a *Process*). The edges of the graph represent (i) dependencies between two artifacts (*wasDerivedFrom*) to track the genealogy of artifacts, (ii) dependencies between two processes (*wasTriggeredBy*) to track the sequence of processes, and (iii) dependencies between artifacts and processes (*used*/*wasGeneratedBy*) to track the consumption and the production of artifacts through processes. Additionally OPM allows to track the links between processes and their enactor agents through *wasControlledBy* dependencies.

However these kinds of dependencies are very generic and are proposed as a basis to track input artifacts and output artifacts produced through process invocations. To distinguish several causal dependencies of the same kind, OPM allows to annotate *used* or *wasGeneratedBy* dependencies with syntactic *roles*. A *Role* is defined in OPM as a particular function of an artifact (or an agent) in a process. The OPM model does not formally define roles but allows to "tag" dependencies between artifacts (or agents) and processes with meaningful labels. In OPM the execution of the sample registration process illustrated in Figure 1 could be translated with these two statements "$Registration_{Process} : used(floating) : Image_{Artifact}$" and "$Registration_{Process} : used(reference) : Atlas_{Artifact}$". The syntactic roles "$floating$" and "$reference$" aim at distinguishing how artifacts are related to processes, but their meaning remains highly dependent on their usage within a given process, and thus, remain highly domain-specific. In the following section we propose to go deeper with the notion of roles, through the proposition of a taxonomy of *Roles* in neuroimaging that aims at (i) enhancing the semantic annotation of services, and (ii) exploiting OPM provenance information to deduce meaningful statements in the context of neuroimaging workflows.

5 ROLE CONCEPTS

To benefit from expert knowledge conceptualized through a domain ontology (such as the OntoNeuroLOG ontology), services involved in e-Science workflows are manually associated to concepts of the ontology. Semantically annotating a service consists in using an ontology to bind technical concepts, *i.e.* elements syntactically describing services, to domain-specific concepts. Most of semantic web-services initiatives, namely OWL-S, WSMO, SWSO, or SAWSDL, distinguish the annotation of the functionality of the service from the annotation of the service parameters which consume or produce data. For instance, let us consider a medical image processing

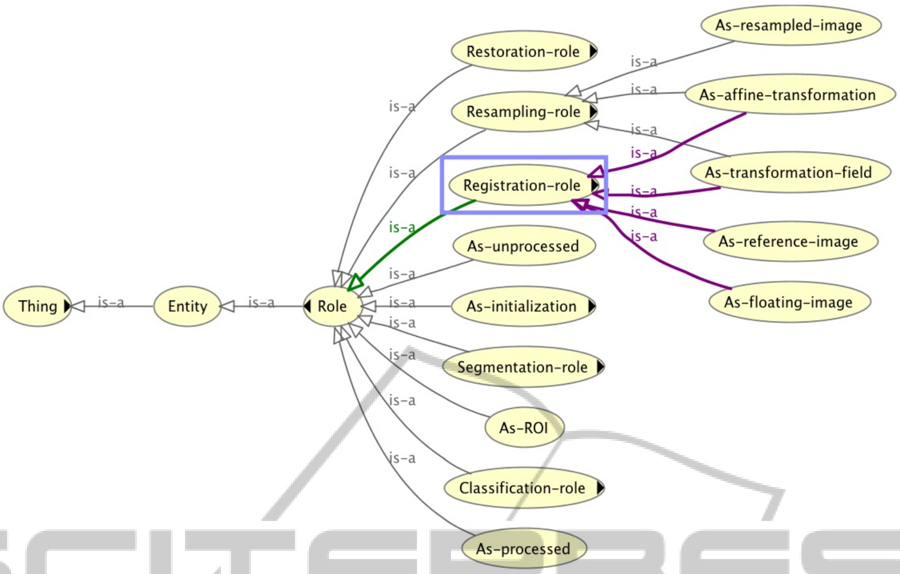

Figure 3: A Role taxonomy characterizing how neuroimaging data can be related to neuroimaging processing tools.

tool performing a de-noising operation. From a technical or syntactical point of view, the service might be implemented by an executable binary taking as input a raw file materializing a noised medical image and producing as output another raw file materializing the resulting de-noised image. From a semantic point of view, this de-noising service might implement a particular kind of algorithm characterizing how the image is processed. This "how" should be described through the annotation of the functionality of the service, *i.e.* a particular class of restoration processing. The service might additionally require a specific medical image format, and a specific modality of acquisition, for instance ultrasound. Moreover, the resulting de-noised image should preserve the input modality; in other words, even de-noised, the image still remains an ultrasound image. The service input/output parameters are usually annotated with concepts describing the nature of consumed or produced data. We will see in the following section that such semantic annotation of the nature of consumed or produced data is often not sufficient to be precisely exploited to produce new domain-specific annotations.

5.1 Differentiating Natural and Role Concepts

Service annotation should also make explicit how consumed or produced data items are related to the processes. For instance, if we consider the registration service involved in the workflow shown in Figure 1, both input parameters should share the same intrinsic nature. Indeed, in this example, the *reference* image parameter and the *floating* image parameter have been acquired both through the same Magnetic Resonance modality (MR) and should be materialized with the same file format. In this geometrical realignment procedure, the two input parameters are not distinguished by their intrinsic nature but rather by their relationship to the registration process, namely *floating* and *reference* images. It is important to note that these two concepts only make sense in the context of a particular kind of image processing, registration. Without the knowledge of "which data is acting as the reference image" or "which data is acting as the floating image", it is difficult to deduce any meaningful information from the execution of the registration workflow, such as "this resulting image can be superimposed with this reference image", or more generally to retrieve images that have been registered with the same reference and thus, that can be superimposed together.

To tackle this issue we propose to distinguish *Natural concepts* and *Role concepts* when annotating semantic service parameters by relying on a domain-specific role taxonomy.

Figure 3 illustrates the taxonomy of roles dedicated to the characterization of the relationships between neuroimaging data and their dedicated processing. *Role concepts* are organized following the main classes of neuroimaging processing like in the OntoNeuroLOG dataset processing ontology.

Relying on this taxonomy of roles, we are now able to precisely annotate the input and output parameters of our image registration service with both *Natural concepts* and *Role concepts*. Both input images are characterized by a same *Natural concept*, T1 weighted magnetic resonance image (T1-MR). T1-MR can be considered as a *Natural concept* because it stands on its own and does not characterize how input data are related with any other entities. On the other hand, service input parameters can be annotated with two distinct *Role concepts* to characterize how input data are related to the registration process. The service input parameter interpreting data as floating (the moving data, that will finally be realigned) is annotated with role *As-floating-image*, and the second service input parameter interpreting data as the geometrical reference is annotated with role *As-reference-image*. Figure 4 illustrates the annotation with *Role concepts* for the two services involved in the full registration use-case workflow.

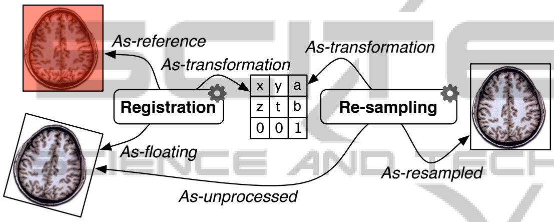

Figure 4: Roles involved in the registration workflow.

Disambiguating the semantic annotation of services, we present in the following section how *Role concepts* are the basis to instrument domain ontologies with reusable inference rules, producing new meaningful statements.

5.2 Integration of OPM and OntoNeuroLOG

The *Roles* taxonomy also acts as a bridge ontology, articulating the two technical ontologies dedicated to the description of services (Web Services) and to the provenance information associated to their invocation (OPM). Indeed, *Role concepts* are associated to the service I/Os (*input/output-variable*) through the same property (*refers-to*) as used to describe the nature of consumed/produced data (OntoNeuroLOG Dataset ontology). Moreover, *Role concepts* are directly extending the OPM *Role* class, so that when recording provenance at workflow runtime, the workflow enactor is able to link *Artifacts* to *Processes* through the newly refined *Roles*.

5.3 Reusable and Service Independent Inference Rules

The use of rule engines (inference engines) is a well adopted data-driven and declarative approach to deduce new conclusions and thus produce new facts from a set of statements. In an OPM-instrumented execution engine, the invocation of services generates provenance statements such as the ones illustrated in Figure 5 for the registration workflow. The graphical syntax introduced by (Moreau et al., 2011) is reused: *Artifacts* are represented by ellipses and *Processes* are represented by rectangles, *used* and *wasGeneratedBy* causal dependencies, parametrized with roles, are represented by plain arrows.

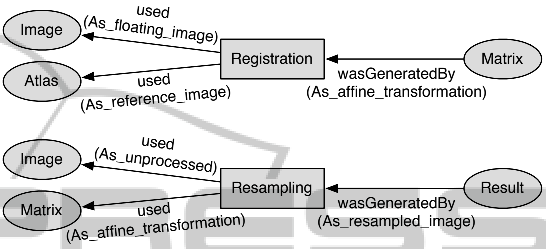

Figure 5: OPM statements recorded through the invocation of the registration workflow.

To automate the production of a statement linking the resulting data to the source data through a domain-specific property, an inference rule is written using the role-parameterized provenance causal dependencies. For instance, Figure 6 illustrates the inference rule deducing the *can_be_superimposed_with* property in the case of the registration workflow. The left part of the implication, the *antecedent* corresponds to the *If* clause of the inference rule and consists in identifying a conjunction of statements necessary to produce the statements expressed in the *consequent*, the right part of the implication (the *Then* clause of the rule). The first two lines assert that processes must refer, for the first one, to a Registration treatment, and for the second one, to a Resampling treatment. In other words, the services invoked by the processes should have been annotated with the corresponding *Natural concepts* of the OntoNeuroLOG domain ontology. The two following lines of the *If* clause identify artifacts and processes through their *Role concepts*: the resulting image is identified through *As-resampled-image*, the registration matrix is identified through *As-affine-transformation*, and the reference image is identified through *As-reference-image*. Finally, when the reference image and the resulting resampled image are identified, the rule engine is able to produce a new statement saying that both images can be superimposed (*can_be_superimposed_with* property of the OntoNeuroLOG ontology).

Using *Role concepts*, domain ontologies can be instrumented with inference rules which remain service independent. Such inference rules can be reused

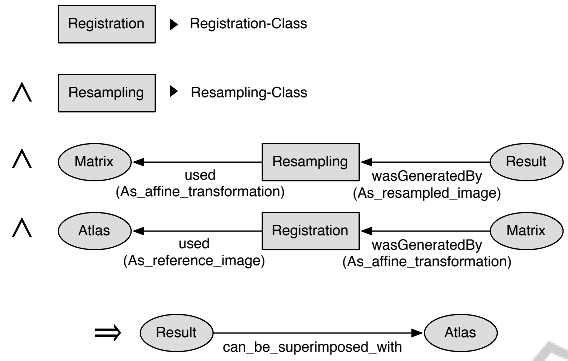

Figure 6: Reusable inference rule automating the annotation of superimposable images.

in the context of several service implementations realizing a same kind of treatment. Let us consider the deployment of a new registration service, implemented with a new algorithm. As soon as this new service is annotated with *Role* and *Natural concepts* of the same class as (or subsumed by) the concepts appearing in the registration inference rule, there is no need to rewrite an inference rule specific to this particular service. As a consequence, workflows involving this new service will also benefit from the generation of annotations stating the “superimposability” of data. With this approach, domain expert can equip their ontologies with inference rules that provide meaningful information to end-users independently from the services deployed. Service providers can focus on their services, transparently reusing such high-level inference rules.

6 IMPLEMENTATION

6.1 System Architecture

Figure 7 schemes the NeuroLOG platform, with a particular focus on its semantic components aiming at enhancing the sharing and enactment of neuroimaging workflows. This deployment shows three collaborating sites A, B, C and end-users interacting with their proper site gateway (Site A) through the client application. Processing tools are syntactically described and instrumented as relocatable bundles through jGASW (Rojas Balderrama et al., 2010) to enable their deployment and invocation on various computing infrastructure. The MOTEUR (Glatard et al., 2008) component enables the design of new experiments as scientific workflows and is responsible for their enactment. The semantic annotation of jGASW services is realized through a dedicated user interface of the client application (*Service annotator*) while the workflow enactor is responsible for recording provenance information at runtime and populating the semantic store with OPM RDF statements. Semantic annotations are managed through local RDF triple stores implemented with the Jena API. The CORESE semantic engine (Corby et al., 2004) is used to perform semantic querying and reasoning over the knowledge base. CORESE is a semantic query/rules engine based on conceptual graphs, supporting RDF(S) entailments and a subset of OWL-Lite entailments: datatypes, transitivity, symmetry and inverse properties.

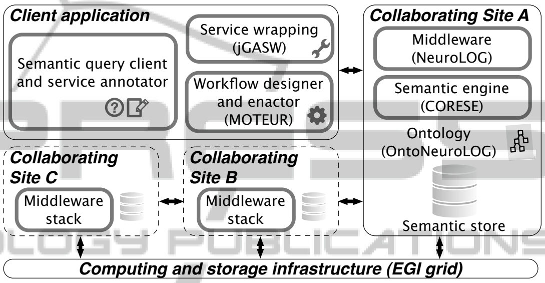

Figure 7: NeuroLOG platform: semantic enhancements to support the sharing and enactment of neuroimaging workflows.

6.2 MOTEUR-S

This section presents the extension of the MOTEUR workflow environment to (i) annotate and catalog jGASW services and (ii) track and query provenance information through the OPM standard. This work is based on the integration of the JSPF plugin framework within MOTEUR, allowing third-party developers to integrate *repository* or *listener* plugins. Repository plugins are dedicated to extending the sources of composable services, and listener plugins (based on the observer design pattern) are dedicated to monitor workflow states at runtime, and trigger specific pre- or post-processing.

6.2.1 Semantically Annotating and Cataloging Processing Tools

The annotation task, generally performed by the processing tool provider, consists in associating to each service port through a dedicated GUI, at maximum one *Natural concept* specifying the semantic nature of the consumed/produced data, and at maximum one *Role concept* characterizing how data is related to the service through this particular port. It is not desirable to associate more than one *Natural concept* or one *Role concept* to a given service port since it will conduct to an ambiguous semantic description of the service. From the service point of view, it would not be possible to determine which *Natural* or *Role concept* characterizes the consumed or produced data. Semantic descriptions can then be saved as a collection of RDF statements, or directly published into the semantic store.

To enhance the overall coherency of workflows at design time, the semantic service catalog can be queried to retrieve services realizing a particular kind of treatment (through an associated *Natural concept* of the domain ontology), or to retrieve services able to consume a particular kind of data at a given step of the workflow construction.

6.2.2 Recording and Querying Provenance Information

When a workflow is started, a new *OPM account* is registered and timestamped. One OPM account is created per workflow invocation, thus easing the retrieval of all provenance annotations generated in the context of a single workflow invocation. For each process invocation, we register an *OPM process* entity, also timestamped, and its consumed and produced data as *OPM artifacts* linked to the *OPM process* through their corresponding causal dependencies *used* or *wasGeneratedBy*. For each causal dependency, we associate an *OPM role* corresponding to the *Role concept* used to annotate the service description. Finally, *OPM processes* are linked through an *OPM wasControlledBy* entity to an *OPM agent*. This agent corresponds to the service description identified by the WSDL URL of the jGASW service deployed. Semantic service description being also identified by the WSDL URL, the system is thus able to retrieve, from an *OPM process* and the semantic catalog of annotated services, all available domain-specific annotations (*Natural* and *Role concepts*) describing the invoked service.

The CORESE semantic engine is used to perform generic queries to retrieve, for instance, from the workflow results, the source data that has been derived through the data analysis workflow. In practice we rely on SPARQL 1.1 property path expressions which provide a compact and powerful language to handle complex graph matching, such as alternative/optional paths, or path length constraints.

6.3 New Knowledge Generation

CORESE provides a forward chaining engine that, from a set of inference rules, saturates its conceptual graph until no new statements can be inferred. Inference rules are expressed through the CORESE rule syntax (non-XML but SPARQL-like), which is very similar to the SWRL proposal from the W3C, also describing an implication between an *antecedent* and a *consequent*. However, CORESE rules differ since they benefit from a limited support of CORESE for OWL-Lite entailments.

In this first implementation, we assume that inference rules instrumenting the domain ontology are provided by the ontology developers and are thus packaged within the domain ontology as complementing files. Inference rules could be applied to extend the knowledge base at any time. Rather than letting the end-user select the suitable inference rule, and trigger the application of the rule, all available rules are systematically applied, triggered by the end of a workflow invocation through a specific MOTEUR event. We consider this automatic application of rules because if the *antecedent* of the rule is not matched, then the rule is not applicable. On the other hand, when an *antecedent* is matched, it makes sense to apply the rule since it has been provided by the ontology developers and has been designed to serve the concerns of the whole user community.

When the MOTEUR listener plugin dedicated to provenance annotations is notified with the end of a workflow invocation, the CORESE semantic engine is populated with (i) the domain ontology (covering both *Nature* and *Role concepts*) and OPM provenance ontology, with (ii) all available inference rules provided with the domain ontology, and (iii) with statements describing the annotated services and OPM statements describing the workflow invocation. Then the forward chaining engine of CORESE is started to produce new inferred statements.

7 CONCLUSIONS AND PERSPECTIVES

E-Science experimental platforms strongly rely on Service Oriented Architectures to assemble flows of data analysis services. However, their usability is hampered by the level of expertise of experiment designers, as they are expected to have a clear understanding of the semantics of the data processing, *i.e.* what kind of data is processed and how they are effectively processed. To improve their usability and assist end-users, ontologies, semantic annotations and reasoning engines are integrated. In this paper, we proposed a clear delineation between *Role* and *Natural concepts* in the domain ontology to disambiguate semantic annotation of service parameters. In addition, the domain ontology can be instrumented with inference rules that leverage the description of *Roles* combined with generic provenance information to enrich our semantic repository with meaningful domain-specific annotations at runtime.

Since this work was implemented in the context of the NeuroLOG experimental platform, the service considered for annotations are web services wrapped with jGASW, a legacy processing tools wrapper targeting large scale distributed infrastructures. However our approach is more widely applicable, and could be implemented using standard web services described through standard service ontologies.

Regarding the sharing of the Role taxonomy of neuroimaging data with other user communities, two approaches could be considered as a continuation of this work: (i) the creation of an OPM profile dedicated to the neuroimaging domain, and (ii) the articulation of the OPM ontology with the DOLCE foundational ontology.

OPM profiles constitute a good opportunity to share knowledge associated to the role of neuroimaging data. Indeed, an OPM neuroimaging profile could be constituted with the two subsets of the OntoNeuroLOG ontology supporting this work, the *Dataset* ontology to extend OPM *Artifacts*, and the *Dataset-processing* ontology to extend OPM *Processes*. The Role taxonomy proposed in this paper could be integrated almost directly.

The second approach, more conceptual, would consist in proposing an OPM ontology whose main classes are grounded to foundational ontologies such as DOLCE or BFO (Basic Formal Ontology). It would allow to smartly articulate OPM and domain ontologies based on foundational ontologies such as BIOTOP (Top-Domain ontology for the life sciences) or OBI (Ontology of Biomedical Investigation) life science ontologies, and thus exploit these ontologies at workflow runtime. Indeed considering our approach from an ontology design perspective, a significant effort is still needed for a complete integration in the OntoNeuroLOG framework, since *Role concepts*, designated through the *refers-to* property, should conform to the DOLCE foundational ontology and its related core ontologies. This ontology integration task could also cover the semantic overlap between OPM *Artifacts* and OntoNeuroLOG *Datasets*. However, in the context of this work, the CORESE semantic engine can still (i) retrieve service description, or provenance statements through SPARQL queries and (ii) produce new meaningful statements through its inference engine.

The concepts developed in this paper are currently being integrated in a prototype platform. In the future, its use in production in the context of the Virtual Imaging Platform (VIP project) will enable the evaluation of our approach. We plan to study the impact of *Role concepts* on four actors in the system: the service providers, the workflow designers, the ontology and inference rule designers, and the final end-users realizing e-Science workflows. Indeed, we plan to measure if *Role concepts* are actually used by service providers to annotate their processing tools, and if they enable to disambiguate service parameter annotations, to finally enable more accurate results when workflow designers query the semantic catalog of services. Moreover, we plan to evaluate if *Role concepts* are actually involved by ontology designers into inference rules to produce new domain specific statements. Finally, we plan to evaluate the production of new annotations at workflow runtime, and its usefulness from the end-user perspective through the analysis of the semantic queries. More precisely, we want to determine if the targets of the semantic queries are annotations inferred from rules involving roles, or if the targets are annotations produced by other means.

Initially applied to computational neurosciences, this work goes beyond this scope, as same principles are planned to be applied in the context of the VIP project, which targets medical image acquisition simulation. It is envisaged to validate the applicability and usability of the delineation of *Role* and *Natural concepts* in domain ontologies to (i) ease the design of simulation workflows (*e.g.* simulated cardiac images through ultrasound modality) and (ii) extend semantic repositories with new meaningful statements describing either simulated data or the simulated organs and their constituting anatomical entities.

ACKNOWLEDGEMENTS

This work was funded by the French National Agency for Research under grants ANR-06-TLOG-024 (NeuroLOG project) and ANR-09-COSI-03 (VIP project).

REFERENCES

Battle, S., Bernstein, A., Boley, H., Grosof, B., Gruninger, M., Hull, R., Kifer, M., Martin, D., McIlraith, S., McGuinness, D., Su, J., and Tabet, S. (2005). Semantic Web Services Ontology (SWSO) [http://www.w3.org/submission/swsf-swso]. http://www.w3.org/Submission/SWSF-SWSO.

Bhagat, J., Tanoh, F., Nzuobontane, E., Laurent, T., Orlowski, J., Roos, M., Wolstencroft, K., Aleksejevs, S., Stevens, R., Pettifer, S., Lopez, R., and Goble, C. A. (2010). BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acid Research*, Article in press.

Corby, O., Dieng-Kuntz, R., and Faron-Zucker, C. (2004). Querying the Semantic Web with Corese Search Engine. In *ECAI*, pages 705–709.

Glatard, T., Montagnat, J., Lingrand, D., and Pennec, X. (2008). Flexible and efficient workflow deployement of data-intensive applications on grids with MOTEUR. *International Journal of High Performance Computing Applications (IJHPCA) IF=1.109 Special issue on Special Issue on Workflows Systems in Grid Environments*, 22(3):347–360.

Gordon, P. M. K. and Sensen, C. W. (2008). Creating Bioinformatics Semantic Web Services from Existing Web Services: A Real-World Application of SAWSDL. In *ICWS*, pages 608–614.

Gruninger, M., Hull, R., and McIlraith, S. (2008). A Short Overview of FLOWS: A First-Order Logic Ontology of Web Services. *IEEE Data Engineering Bulletin.*, 31(3):3–7.

Henriksson, J., Pradel, M., Zschaler, S., and Pan, J. Z. (2008). Ontology Design and Reuse with Conceptual Roles. In *Proceedings of the 2nd International Conference on Web Reasoning and Rule Systems*, RR '08, pages 104–118, Berlin, Heidelberg. Springer-Verlag.

Kassel, G. (2005). Integration of the DOLCE top-level ontology into the OntoSpec methodology. Technical Report 2005-10-18, University of Picardie, Amiens, France.

Kopecký, J., Vitvar, T., Bournez, C., and Farrell, J. (2007). SAWSDL: Semantic Annotations for WSDL and XML Schema. *IEEE Internet Computing*, 11(6):60–67.

Lord, P., Alper, P., Wroe, C., and Goble, C. (2005). Feta: A Light-Weight Architecture for User Oriented Semantic Service Discovery. In *European Semantic Web Conference*, pages 17–31. Springer Berlin / Heidelberg.

Martin, D., Burstein, M., McDermott, D., McIlraith, S., Paolucci, M., Sycara, K., McGuinness, D. L., Sirin, E., and Srinivasan, N. (2007). Bringing Semantics to Web Services with OWL-S. *World Wide Web*, 10(3):243–277.

Montagnat, J., Gaignard, A., Lingrand, D., Rojas Balderrama, J., Collet, P., and Lahire, P. (2008). NeuroLOG: a community-driven middleware design. In *HealthGrid*, pages 49–58, Chicago. IOS Press.

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

Rojas Balderrama, J., Montagnat, J., and Lingrand, D. (2010). jGASW: A Service-Oriented Framework Supporting High Throughput Computing and Non-functional Concerns. In *IEEE International Conference on Web Services(ICWS 2010)*, Miami, FL, USA. IEEE Computer Society.

Roman, D., de Bruijn, J., Mocan, A., Lausen, H., Domingue, J., Bussler, C., and Fensel, D. (2006). WWW: WSMO, WSML, and WSMX in a Nutshell. *The Semantic Web – ASWC 2006*, pages 516–522.

Sheth, A. P., Gomadam, K., and Ranabahu, A. (2008). Semantics enhanced Services: METEOR-S, SAWSDL and SA-REST. *IEEE Data Eng. Bull.*, 31(3):8–12.

Sowa, J. F. (1984). *Conceptual structures: information processing in mind and machine*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

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

Vitvar, T., Kopecky, J., Viskova, J., and Fensel, D. (2008). WSMO-Lite Annotations for Web Services. In *5th European Semantic Web Conference (ESWC2008)*, pages 674–689.

Withers, D., Kawas, E., McCarthy, L., Vandervalk, B., and Wilkinson, M. (2010). Semantically-guided workflow construction in Taverna: the SADI and BioMoby plug-ins. In *Proceedings of the 4th international conference on Leveraging applications of formal methods, verification, and validation - Volume Part I*, ISoLA'10, pages 301–312, Berlin, Heidelberg. Springer-Verlag.

Wolstencroft, K., Alper, P., Hull, D., Wroe, C., Lord, P. W., Stevens, R. D., and Goble, C. A. (2007). The myGrid ontology: bioinformatics service discovery. *IJBRA*, 3(3):303–325.