SEMANTIC INDEXATION OF WEB SERVICES FOR COLLABORATIVE EXPERT ACTIVITIES

Jérôme Dantan
LAMSAD Laboratory, Esitpa-APCA / CEDRIC Laboratory, CNAM
3 rue du Tronquet CS 40118 76134 Mont St Aignan Cedex, France / 292 rue Saint-Martin 75003 Paris, France

Yann Pollet
CEDRIC Laboratory, CNAM
292 rue Saint-Martin 75003 Paris, France

Salima Taibi
LAMSAD Laboratory, Esitpa-APCA
3 rue du Tronquet CS 40118 76134 Mont St Aignan Cedex, France

ABSTRACT

In many research domains, scientists have to cope with large amounts of data issued from multiple sources, looking for regular patterns in raw data that may suggest models. The extraction of knowledge from data in general is widely used in fields such as biology, economics, etc. For this, they develop models in their respective area of research. The efficiency of the experts depend on their ability to cooperate: they share data and models. Indeed, each domain is both consumer of treatments provided by another team, and provider for the benefit of other areas. Each team of experts have to make available and publish business processes for the benefit of other teams, that allows them to have the most up to date treatments. In this work, we propose a work environment based on semantic web services indexation whose benefits will offer (1) a platform to publish and reuse treatments, accessible through the Internet, (2) services composition to chain treatments and assisted and/or automated selection of useful services. For this, we define two kinds of services: knowledge services, that provide business processes and data services that are wrappers that provide the relevant data from heterogeneous data sources. They will be accompanied with an existing OWL ontology, used by the experts of the considered domain. As a result, experts will be able to operate and cooperate through heterogeneous application environments and all around the world. An example on modeling farm durability indicators will be presented.

KEYWORDS

Collaborative platform, interoperability, knowledge, model, ontology, reusability, semantic indexation, services orchestration.

1. INTRODUCTION

Scientists from many research domains work collaboratively; they produce and exchange data, partial results and models on that other experts rely to build new results. Most studies need to be addressed in a multidisciplinary manner and each expert deals with a particular aspect of the problem. It may be useful for experts to take into account new data, share partial results, update their own indicators and models and to take advantage of new hardware platforms, new measures and updated indexes in real time, with the objective of pooling resources. That's why a collaborative process is needed. For reasons of efficiency, confidentiality and heterogeneous environments, the source code of processing functions, that contain complex business processes, cannot be directly available. In addition, if any expert downloads data on his local computer, he may forget to update it regularly, that implies risks of errors and outdated data. So each expert can only make available the entry points for remote calls of services, and not the entire codes and data.

In this context, the Semantic Web approaches look promising, especially those related to ontologies and services orchestration. We have designed a system that offers a mean to publish models that experts have developed, with pertinent indexation, assisted or automated identification of the useful services, and finally their composition, that allows the execution of processing chains. The services must be semantically indexed to assist or even automate their reuse through the Web, using ontologies of the considered domain. The rest of the paper is organized as follows. Section 2 presents an analysis of the agriculture domain. Section 3 presents the state of the art. Then section 4 presents the approach we propose. Next section 5 presents the semantic model, including the services indexation, the algebra we propose and a use case on modeling farm durability indicators. Section 6 illustrates the design of the system. Finally, section 7 presents our conclusions and perspectives for future work.

2. EXPERT ACTIVITIES: THE AGRICULTURE DOMAIN EXAMPLE

Nowadays, the evaluation of human activities sustainability is a major preoccupation for scientists, e.g. experts in the domain of agriculture not only evaluate farms efficiency and production, but also their sustainability. The evaluations are performed thanks to models and indexes that are goals to reach. They need a multidisciplinary expertise e.g. the domain of agriculture implies many areas: agronomy, ecology, rural economy, etc. Rural economy experts focus on economic durability and performances of farms according to both economic parameters (income, number of employees, ...) and environmental parameters (soil quality indexes, farming practices, ...). Conversely, agronomy experts are interested in the sustainability of a farm with the economic performance of farms. Consequently, each model is both consumer of treatments provided by other domains and provider for the benefit of other domains.

A goal is achieved when the more specific sub-goals are achieved e.g. the agro-ecological sustainability assessment is based on both biological assessments such as biodiversity, soil quality, etc and area organization parameters such as crop rotation, management of forage areas, etc, that are goals at a more specific level. At the below, the goals are simply collections of raw data. Researchers compute raw data from many sources, both local and from the Internet that cannot be used directly, e.g. relational databases and complex business processes. Consequently, many computer programs and SQL queries have been developed for the data to be interpreted by other software. However, they are specific to one data source. Whether to extract data from other sources, they need to reuse the code already written.

In the model we propose, each treatment is considered as a class of goals instance. Indeed, the researchers propose the indicators they have conceived of the test data sets and publish them, considering that (1) indicators are designed by the researcher from the aggregation of other indicators developed by other researchers, and (2) computations on data sets may be created by complex processes as they may be based on performances of remote services.

3. RELATED WORK

Ontologies are widely used to capture and organize knowledge about a particular domain. They are used as an index to retrieve specific data (Garcia, R. and Celma, O., 2005), to infer new knowledge (World Wide Web Consortium (W3C), 2004), to semantically annotate multimedia data (Castano, S. et al, 2007), to find out Web Services automatically (Martin, D et al, 2007), or to match knowledge with other knowledge for a more general purpose (Cruz, C. and Nicolle, C., 2011). Ontologies offer significant benefits to collaborative service-oriented systems, such as interoperability and reusabilility. We will use Web Ontological Language (OWL) related technologies that provide model flexibility, explicit representation of semantics, out-of-thebox reasoners and proliferation of freely available background knowledge (Castellani, S et al, 2011), e.g. the AGROVOC OWL ontology (FAO, 2011).

In addition, the KDD reflects the whole process from extraction of raw data, pre-treatment to their interpretation by the scientist, as illustrated by figure 1.

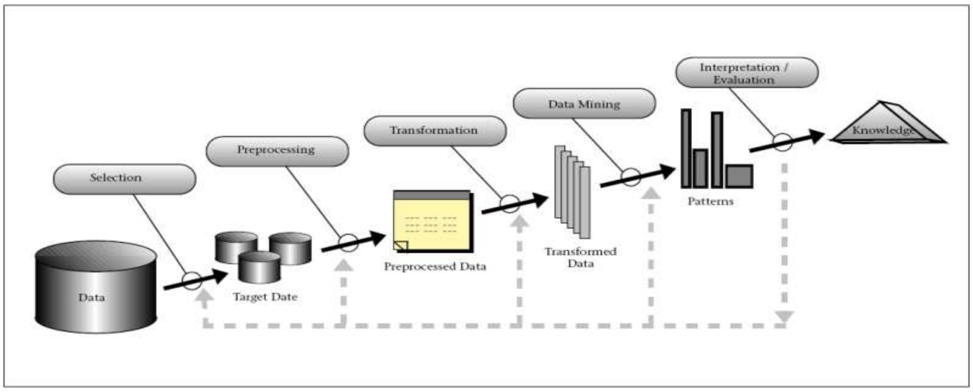

Figure 1. An overview of the steps that compose the KDD process (Fayyad, U. et al, 1996).

Some prototypes of collaborative platforms already exist. Nevertheless, no platform in the agriculture domain offer a mean to publish the treatments experts have developed, with pertinent indexation, assisted or automated identification of the useful services, and their composition that allows the execution of processing chains, with services that are semantically indexed to assist or even automate their reuse through the Web, using ontologies of the considered domain. Consequently, the problem is considered here with a completely different scale. Indeed, the system we propose provides interoperability, reusability and new knowledge inference thanks to our use of both ontologies and services orchestration.

4. PROPOSED APPROACH

The approach we propose may provide interoperability with other systems in heterogeneous environments. The services will be semantically indexed to perform their reuse and infer new knowledge. They are not limited to standard statistical treatments, but include complex business processes that will be indexed and chained thanks to OWL related technologies, so that it is neither possible to copy and run it on his machine, nor possible to include in a package. The users are involved in the loop to improve the process on the Internet and they interact with the system by submitting their models, indicators and data.

In the platform we have designed, the operation of matching compares the abstract services that are defined by the user with the existing concrete services, to ensure their strict equivalence, isolate their common elements or enumerate their differences. This operation therefore consists of three distinct primitives: (1) testing the equivalence of two services (Equivalence(Si, Sj)); (2) calculating the intersection of two services (Match(Si, Sj)) i.e. the set that contains all elements of S(i) (resp. S(j)) that also belong to S(j) (resp. S(i)), but no other elements; (3) calculating differences between two services (Mismatch(Si, Sj)) i.e. the set of elements in S(j), but not in S(i). We only consider the principle of exact matching to ensure that a service meets a need expressed by the user because, in the system we propose, the various input parameters and output of a service have accurate meanings that require exact formulations. The collaborative platform (1) will offer an orchestration of concrete services by searching in the services databases for those that are appropriate and (2) will be responsible for executing the orchestration chosen by the user.

We need to index the services of the collaborative platform. The services registery system must remember, for a given service, the triggering events, the set of actions a service can perform, events that affect an agent's course of action, commitment strategy and conflict resolution strategy (Quynh-Nhu Numi Tran and Graham Low, 2008) e.g. initial state, commitment strategy, actions, in parameters, out parameters. Above all, the services registry must remember the purpose for that a given service has been designed, that gives the sense of the service *ie* it must perform a semantic capture, e.g. to evaluate indicators, models, to format data, etc. To describe the system, we have defined a knowledge model with three different layers, as illustrated in figure 2.

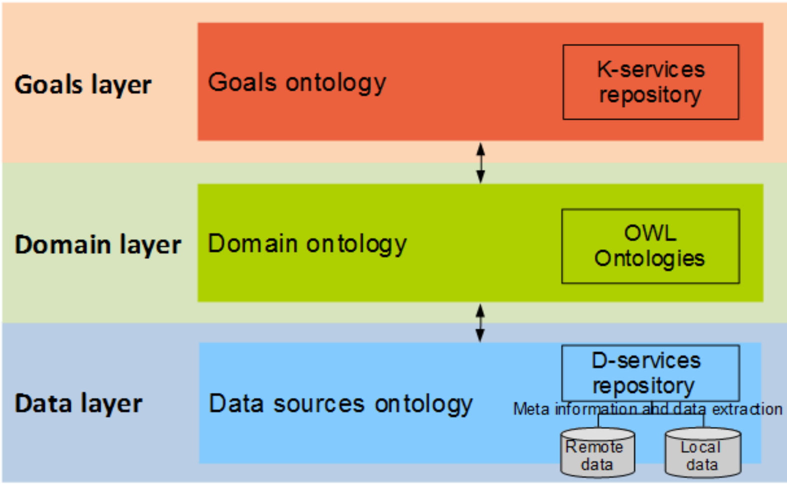

Figure 2. The knowledge model.

(1) The domain layer contains the concrete concepts that will be useful to be considered for experts e.g. fields, microbiological elements, etc, that are typical examples of characters that may have a role in evaluating the durability of farms. For this, we use a widely used ontology of the domain. The Food and Agriculture Organization of the United Nations (FAO) provides a multilingual agricultural ontology called AGROVOC (W3C, 2004).

(2) The data layer provides not only the data sources such as database schema, data warehouses, database systems, etc, but also the computer programs (wrappers) have been coded by software developers and that provide the aggregated relevant data. In the model we propose, the relevant data are provided by Data-services (D-services) that are expressed thanks to the elements of the domain ontology.

(3) The goals layer contains the ontology of the elements that we have to evaluate. They are hierarchically organized e.g. on the top of the hierarchy stands the sustainability of the considered farm; at a more specialized point of view, the domain-oriented indexes like economic sustainability and environmental sustainability. The more specific the elements are, the more detailed the indexes are. At the bottom of the goal layers stand the goals that consist in gathering the relevant data. Every new model and program proposed by the researcher become a new goal instance belonging to our system. The inputs and outputs of the goals are not necessarily elements of the ontology, but they are more complex data. In this paper, the services that provide the goals to reach are called Knowledge Services (K-services). In the next section, we present the semantic model provided by K-services and D-services.

5. THE SEMANTIC MODEL

We define new kinds of classes properties, as an extension of OWL, to index the services.

5.1 Data-services indexation

On the one hand, data processing experts create D-services, e.g. SQL queries and more complex programs such as computations of indicators. A D-service is characterized by the description of the entity it issues. This is done by adding the notion of derived attribute: the derived property that results in a reference path, deduced from other classes thanks to a transitive link and OWL operators.

5.2 Knowledge-services indexation

On the other hand, experts of the considered domain create their own K-services to enrich the collaborative platform. A K-service is defined by the types of input and output data, the preconditions to validate the call and the goal that the K-service provides. The inputs and outputs may be elements of the ontology or more complex data. They may be in particular:

- Partial descriptions of objects, e.g. inputs such as the age and marital status of a farmer. Indeed, a farmer is described by many things in more than age and marital status, but these two elements are datatype properties characterizing an object belonging to the "farmer" class. a farmer is reduced to a couple (age, marital status). This is analogous to a relational projection.
- Aggregated descriptions of objects, e.g. the cattle of the farm does not come directly from the farmer object, but was extracted by considering other objects in connection with the farmer object, e.g. the class “farmer” has an object property “farm”. Nevertheless, the cattle belongs to the farmer.
- Tuples data.

A K-service is characterized by the purpose it instantiates, added to an OCL-style (Object Constraint Language) pre-condition (if any), e.g. (`farm:Cattle=“Cow” AND Culture=“Wheat” and (Region=“Normandy” OR Region=“Brittany”`). It needs data and to activate other goals at more specific levels to deliver the final result. We do not express its elements by considering the data types only, but also the ontological model that acts as a reference e.g. the inputs of a service are classes from the ontology of the domain or goals/subgoals from the ontology of goals. Goals are related to domain ontology classes e.g. the agro-ecologic durability is evaluated on a farm (and not on cows or farmers...). To deal with it, we define the notion of factor attribute that is a goal.

5.3 Algebra

OWL distinguishes between two main categories of properties that an ontology builder may want to define:

- Datatype properties for which the value is a data literal.
- Object properties for which the value is an individual.

In this section, we present our services indexation model from a theoretical point of view i.e. the algebra that formalizes the notions of derived and factor properties, as an extension of OWL.

OWL distinguishes six types of class descriptions: a class identifier (a URI reference); an exhaustive enumeration of individuals that together form the instances of a class; a property restriction; the intersection of two or more class descriptions; the union of two or more class descriptions; the complement of a class description.

Let us define the operators that extend the basic OWL axioms:

- Definition 1: A derived property P attached to a class C is a property deduced from other classes thanks to transitive links that may be chained thanks to the basic OWL operators such as navigation, restriction, and value, added to an additional aggregation operator, such as Min, Max, Sum, Median, Average or Standard Deviation, that associate the values of entities to the result of the operator, applied to their input values.
- Definition 2: A factor property P attached to a class C is a derived property that belongs to the ontology of the goals, defined thanks the complex business processes we defined above and that extend the notion of property of OWL.
- Definition 3: The whole of the generalized properties of the class C are formed of the whole of its natives properties added the whole of its derived properties.
- Definition 4: A description Descr(C) of the C class is a a new class C', the properties of that are a subset of the generalized properties of the C class e.g. the class (farmer.age, farmer.sex,

cultivatedSurface, income) is a description of the Farm class. This is analogous to an object projection and implies a new OWL class equivalent to the C class. The C class itself is one of its own descriptions. The notion of description characterizes a data source, by the selection of the pertinent data. So, the description defines a class C' that contains exactly the same instances as the previous class C and matches to the OWL notion of equivalent class.

5.4 Examples

The farm class has native properties, derived properties and factor properties that we have written in OCL.

5.4.1 Native properties

Data properties (e.g. sales) and object properties (e.g. the farmer who is a “Human” object; the parcels that are “Parcels” objects) are native properties of the “farm” class.

5.4.2 Derived properties

The “farmerAge” and “surface" properties respectively correspond to “farmer.age” and “farm.SUM(Parcels.Surface)”.

The “wheatCultivatedSurfaceArea” property corresponds to the SUM of the cultivated surface area of each parcels that are associated to the “CultivatedParcel” subclass and the “Wheat” culture type:

```
WheatCultivatedSurfaceArea =
SUM(NAVIGATION(Farm.Parcel).SELECT(Parcel.subclass=”CultivatedParcel”
                                                           AND CultureType=”Wheat”))
```

5.4.3 Factor properties (goals)

The IDEA method is an aggregated model to assess farms sustainability (Vilain. L. et al, 2008). It includes seventeen agro-ecological goals that allow a good economic efficiency for an environmental cost as low as possible e.g. biodiversity, space organization, animal welfare, quality of products, ethics, employment, etc. Suppose that a researcher has developed an aggregated soil quality index that can be added to the "space organization" goal of the IDEA model. The researcher may automatically assess the effects of his new indicator, apply it to other farms by querying other databases and share it with other experts. A canvas of the “soil quality index” goal classes is illustrated in figure 3.

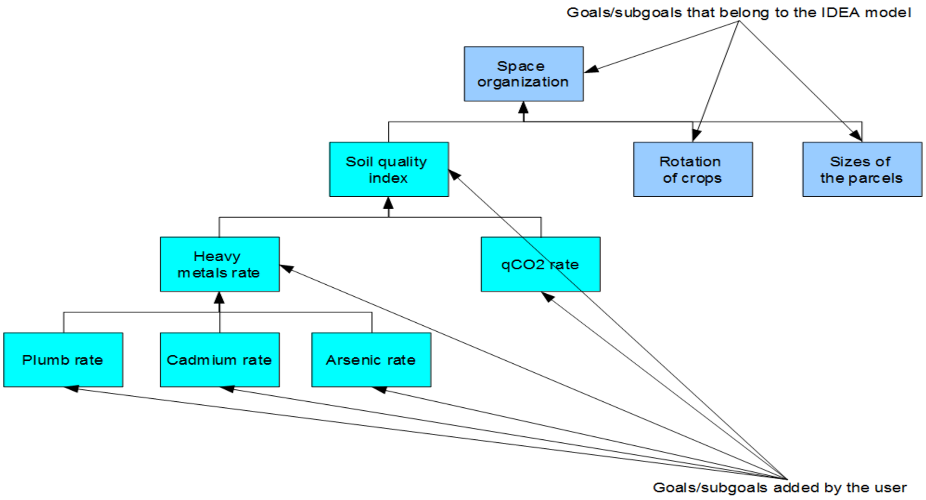

Figure 3. A canvas of goal classes

The inputs of the "soil quality index" Knowledge-service are the rates of heavy metals and respiration to microbial biomass ratio (qCO2). The output is a "soil quality index" object belonging to the goals ontology:

```
SoilQualityIndex = SUM(Farm.qCO2Rate.MEDIAN, Farm.HeavyMetalRate.MEDIAN)
```

with

```
Farm.HeavyMetalRate.MEDIAN = SUM(Farm.CadmiumRate.MEDIAN,
                                 Farm.PlumbRate.MEDIAN,
                                 Farm.ArsenicRate.MEDIAN)
```

with

```
Farm.CadmiumRate = SUM(Farm.Parcels.CadmiumRate)
Farm.PlumbRate = SUM(Farm.Parcels.PlumbRate)
Farm.ArsenicRate = SUM(Farm.Parcels.ArsenicRate)
```

As described above, new elements of the ontology of goals can be added to create or modify sustainability models such as the IDEA model.

6. THE COLLABORATIVE PLATFORM

The figure 4 illustrates the components of the future collaborative platform.

- The ontologies management layer provides one or more ontologies of the considered domain.
- The services management layer provides the K-services and D-services indexation systems and registries and the orchestration of the services.

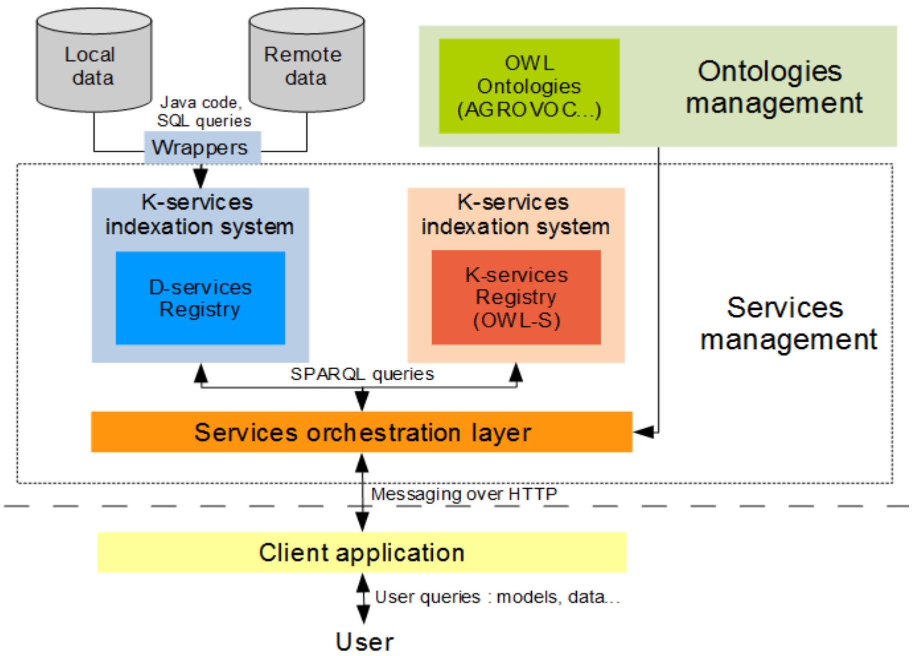

Figure 4. Components of the future collaborative platform.

The development of our collaborative platform is in progress. The figure 4 introduces the proposed scenario:

1. Request sent by the user via the the communication interface to the service orchestration layer. The client application offers a mechanism for the formulation of new goals.
2. Search for the existing services that match to the user request in the services registries.
3. If no matching services, proposal to create new services to the user.
4. If matching services, services composition and then sending results to the user.

7. CONCLUSION

As a conclusion, we provide a way to index services designed by experts thanks to an ontology of goals, an ontological formulation that formalizes the data, treatments, and goals developed by computer scientists, and ad hoc notions of aggregated and description properties that extend OWL. The functions we have described through this formalism have been validated in the field of agronomy, and the results are encouraging. However, the model applies more generally to many areas and we have a good basis to evaluate, refine and evolve our platform with experts from different domains, on other situations of use. Furthermore, much work remains to be done, including the orchestration of the services, a query language for defining abstract processes and the declarative language that will define the objectives of services.

REFERENCES

Castano, S. et al, 2007. Ontology Dynamics with Multimedia Information: The BOEMIE Evolution Methodology. *Proceedings of International Workshop on Ontology Dynamics (IWOD) ESWC 2007 Workshop.* Innsbruck, Austria.

Castellani, S et al, 2011. A knownledge-based system to support legal case construction. *Proceedings of Knowledge Engineering and Ontology Development conference.* Paris, France, pp. 15-27.

Cruz, C. and Nicolle, C., 2011. A graph-based tool for the translation of xml data to OWL-DL ontologies. *Proceedings of Knowledge Engineering and Ontology Development (KEOD) IC3K 2011 conference.* Paris, France, pp. 361-364.

Fayyad, U. et al, 1996. From Data Mining to Knowledge Discovery in Databases. *AI Magazine,* Vol. 17, No. 3, Fall, pp 37-54.

Food and Agriculture Organization of the United Nations (FAO), 2011. *AGROVOC.* http://aims.fao.org/website/About/sub.

Garcia, R. and Celma, O.: Semantic Integration and Retrieval of Multimedia Metadata, *Proceedings of 4rd International Semantic Web Conference*, Galway, Ireland (2005).

Martin, D et al, 2007. Towards Semantic Annotations of Web Services: OWL-S from the SAWSDL Perspective. *Proceedings of OWL-S Experiences and Future Developments Workshop at ESWC 2007.* Innsbruck, Austria.

Numi Tran, Q and Low, G., 2008. MOBMAS: A methodology for ontology-based multi-agent systems development. *Information and Software Technology,* Vol. 50, pp 697-722.

Vilain. L. et al, 2008. *La méthode IDEA Indicateur de durabilité des exploitations agricoles.* Educagri, France.

World Wide Web Consortium (W3C), 2004. *OWL Web Ontology Language reference. http://www.w3.org/TR/owl-ref/.*