Integration of a Domain Ontology in e-Science with a Provenance Model for Semantic Provenance Generation in the Scientific Images Analysis

Lucélia de Souza[1,2], Maria Salete Marcon Gomes Vaz[2,3]

[1]Departament of Computer Science – University of Western of Parana (UNICENTRO)
Rua Camargo Varela de Sá, 03, CEP 85040-080 – Guarapuava/PR – Brazil

[2]Department of Informatics – Federal University of Parana (UFPR) Rua Cel. F. H. dos Santos, 100, CEP 81.531-980 – Curitiba/PR - Brazil

[3]Department of Informatics – State University of Ponta Grossa (UEPG) Av. Carlos Cavalcanti, 4748, CEP 84.030-900 - Ponta Grossa/PR - Brazil

{lucelias,salete}@inf.ufpr.br

Abstract. This paper describes the integration of a domain ontology in e-Science with a provenance model for semantic provenance generation in the scientific images analysis. The domain ontology is related with images obtained from the CoRoT Telescope, where the exoplanets search require detrend algorithms as preprocessing, improving the chance to detect planetary transits. In order to retrieve standardized information regarding the origin and facilitate the monitoring of information, the Proof Markup Language – PML was chosen as provenance common model due to its characteristics of modularity, reuse, interoperability and the possibility of justify how conclusions were obtained. As contribution of this paper, the integration of the ontologies presented enables getting information from the domain and justify the conclusions, through a standardized provenance model, allowing logical inference and semantic interoperability.

1. Introduction

In the scientific images analysis, information of provenance provide the source of processing, allowing share, reuse, reprocessing and do further analysis in data and process. The semantic provenance [Sahoo et al 2008] is related with the Semantic Web and can be obtained by means of ontologies [Borst 1997], which represent the knowledge, structuring information in an organized manner and generating semantic in the data.

In this work, a domain ontology was developed in the Ontology Web Language – OWL2 called CorotDataAnalysisOntology (crtdao) [de Souza et al 2011] allowing to extract domain information in the scientific images analysis. It relates with images obtained from CoRoT Telescope[1], which provides thousands of light curves in format Flexible Image Transport System – FITS [Hanisch et al 2001]. In the analysis of these images, the search of planets outside of Solar System (exoplanets) requires detrend

[1] CoRoT Archive: http://idoc-corotn2-public.ias.u-psud.fr/

and/or filter algorithms as preprocessing for removing phenomena that may occur suddenly, such as random jumps and/or trends, slow and gradual changes in certain properties of the images, under whole range of the investigation. So, different detrend and/or filter algorithms can be applied to treatment of these phenomena, improving the chance to detect planetary transits (Figure1).

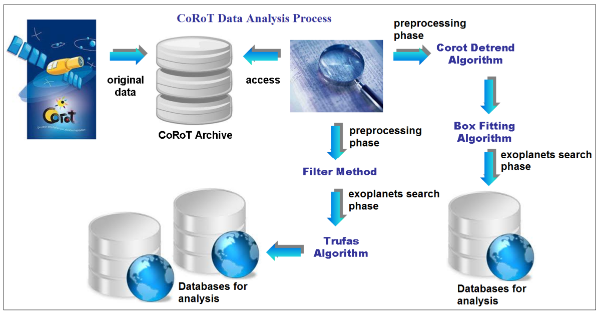


Figure1. CoRoT Data Analysis Process

However, domain ontology can be insufficient, semantically, for generation and sharing of provenance information. It is necessary to make use of a common model for the provenance generation as means to allow interoperability, reuse and extension of ontologies [McGuinness et al 2007]. In this case, provenance models can be integrated with domain ontologies because their use increases the understanding of users about how answers were generated and also facilitates the acceptance of the results. Among the provenance models existing, Provenir [Sahoo et al 2009], Open Provenance Model – OPM [Moreau et al 2011] and Proof Markup Language – PML [McGuinness et al 2007] stands out, being workflow-based systems. These models were analyzed for use in the scientific images analysis, being chosen the PML due to its characteristics of modularity, reuse, interoperability and mainly by allow us to justify how were obtained the conclusions.

The objective of this paper is to present the integration of the domain ontology with the PML Provenance Model, contributing to enrich the scientific images analysis with semantic and standardization. This integration enables getting the information from the domain and justifies the conclusions, by means of inference steps involving which the inference engine, inference rule and/or source used to generate it.

This paper is structured as follows, besides this introductory section. The second section describes about data provenance and workflows. The third section describes about domain ontologies and provenance models and the next section presents their integration. The fifth section brings the related works, followed of the conclusions and the future works.

2. Data Provenance and Workflows

Provenance means origin or source. In the scientific images analysis, provenance information proves the correctness of the resulting data, being regarded by Tan (2007) as important as the result itself.

The provenance information can have granularity in the fine-grain and coarse-grain forms [Tan 2007]. The first form involves the data derivation and storage in databases how proposed by authors Tan (2007), Buneman et al (2007), Cheney et al (2009), among others. There are two approaches, such as metadata annotation and nonannotation approach, through queries and inverse functions used for data transformations. The second form involves activities and processes used to perform tasks of complex scientific data by mean of scientific workflows [Davidson and Freire 2008], where can be human interactions during the execution of processes flow.

2.1 Semantic Provenance in the Scientific Images Analysis

This paper stands out by enriching the data analysis semantically, related to FITS images that are available in the CoRoT Archive to exoplanets search. During the execution of detrend and/or filter algorithms, provenance information can be stored in the FITS images header.

The FITS Standard [Hanisch et al 2001] establishes rules for use of these images, which differs of the traditional format of images, due to its basic structure formed by a header containing metadata (data about data) such as SIMPLE, BITPIX, COMMENT, HISTORY, among others, and a matrix used for storing binary data.

However, the FITS specification does not contemplate the addition of provenance metadata, describing the use of HISTORY metadata to store steps executed. This form of provenance generation is free text, not being machine readable, impeding its use by software agents.

In scientific images analysis, provenance information records steps performed and generating knowledge in order to avoid reprocessing and contribute to sharing, reuse and analysis further. So, the metadata storing in images header or in the databases is insufficient, semantically, to generate provenance. This information is useful for local researchers, but not enough to share, reuse and reprocessing by scientific community. There is a need for standardization of provenance metadata to be generated and stored, just as it takes more detailed information to contemplate the real needs of researchers as to the semantic knowledge about the data generation over time. Accordingly, the next section describes the development of domain ontology in this environment.

2.2 Domain Ontology in e-Science

Ontology is defined as a formal and explicit specification of a shared conceptualization [Borst 1997]. It is characterized as a mean of representing knowledge, structuring information in an organized manner of a domain and generating semantics in the data.

The development process of the domain ontology proposed is based on Ontology Development 101 [Noy and McGuinness 2001]. We started by identifying a set of competency questions from domain that must be answered by the ontology, such as: *What are the statistical techniques (Linear, Polynomial, among others) used by detrend algorithms?; The CoRoT Detrend Algorithm treats which systematic effects?; What transit algorithm had the type of method Least-Squares?;* among others. From these questions, were identified classes, their relationships and the instances. Restrictions are declared using axioms and/or rules, providing semantics and allowing inferences.

The Protégé 4.1 tool [Knublauch et al 2004] was used to the development of the domain ontology and the generation of knowledge base. The used language is OWL 2.0, recommend by World Wide Web Consortium - W3C and based on Descriptive Logic – DL [Baader 2003]. Pellet 2.2 is used to verify its consistency.

An OWL ontology in abstract syntax contains annotations, axioms and facts. However, the use only of axioms presents expressive limitations, mainly with the use of properties such as composition of roles [Horrocks et al 2005].

The composition of roles as '*isAlgorithmDetrendPolynomialOf*' shows an example. If an algorithm is *AlgorithmDetrend* and the method type is *Polynomial*, then the algorithm is an *AlgoritmDetrend* of the *Polynomial* type. The relationship between the composition of the '*isAlgorithmDetrendOf*' and '*isMethodTypePolynomialOf*' properties and the '*isAlgorithmDetrendPolynomialOf*' property is limited to the form $P \circ Q \sqsubseteq P$, in order to maintain decidability. The composition of two properties is a subproperty of one of the composed properties, that is, the complex relationship between composed properties cannot be captured. This is the case of '*isAlgorithmDetrendPolynomialOf*' property that cannot be captured because it is not one of '*isAlgorithmDetrendOf*' not '*isMethodTypePolynomialOf*'.

So, the complex axiom '*isAlgorithmDetrendOf*' $\circ$ '*isMethodTypePolynomialOf*' $\sqsubseteq$ '*isAlgorithmDetrendPolynomialOf*' presents the form $R \circ S \sqsubseteq T$ and $T \circ S \sqsubseteq R$, because exists cyclical dependences in the definition, violating the irreflexivity. This verification is important in relation of the decidability, because such cyclical dependences can induce undecidibility and the use in an ontology should be restricted. One way to address this problem is extend OWL with a more powerful language to describe properties.

Horrocks et al (2005) extends axioms OWL DL to allow rule axioms (a Semantic Web Rule Language - SWRL), in the form: *axiom::=rule*. In the human readable syntax, a rule has the form antecedent→consequent, an implication between an antecedent (body) and consequent (head).

Informally, a rule means "if the antecedent hold (is true), then consequent must also hold". The antecedent and consequent of a rule consist of zero or more atoms, which can be of the form *C(x)*, *P(x,y)*, *sameAs(x,y)* or *differentFrom(x,y)*, where *C* is an OWL DL description, *P* is an OWL property, *x* and *y* are either variables, individuals or data values. Multiple atoms in an antecedent are treated as conjunction and multiple atoms in a consequent are treated as separate consequences [Horrocks et al 2005].

The Protégé 4.1 Tool allows working with rules from View Rules. Pellet supports reasoning with SWRL rules, which interprets SWRL using the DL-Safe Rules notion, where rules will be applied only to named individuals in the ontology.

2.3 Analysis of the domain ontology as to semantic integration

The domain ontology was evaluated by domain experts as the terms used as well by ontologists. Also was formalized in an extension of the DL called *SROIQ* [Horrocks et al 2006], which presents characteristics of the expressiveness, decidability and robust computational properties, being an extension more expressive than the Attributive Language, the most basic family of DL.

The formalization allow us to specify the ontology independently of the domain, contributing to verification and validation of axioms assertional, terminological and role inclusion used, well as allows to infer knowledge. With the formalization in *SROIQ* DL, under OWL 2, recommended in 2009 by the W3C, also is possible to verify the consistency of the knowledge base. In this way, it is feasible to enrich the scientific images analysis with semantic and standardization.

The domain ontology proposed in [de Souza 2011] presents as main classes: *DataSet*, *Methods*, *Technique*, *AlgorithmBase*, *PeriodicSignalShape*, *MethodType*, *Algorithm*, *Software*, *Metadata*, *Person*, *Run*, *Telescope*, *Language* and *SistematicEffectType*. *Header* class was created to relate header specific metadata of FITS images and the *Database* class, related with the storage location. The Language class was specified in *ProgramationLanguage*. However, aiming semantic integration in e-Science, a domain ontology developed was evaluated in relation to existing ontologies (Figure2).

The VSTO ontology[2] stands out as an ontology open-source, extensible and reusable in the area of solar-terrestrial physics, which supports interdisciplinary projects of virtual data collections. This ontology was analyzed, and made the following adjustments in the domain ontology: i. *Telescope* class was inserted as a subclass of *Instrument* and were also imported from VSTO the following classes: *DataProduct* related with FITS images, which were previously represented as *DataSet*; *vsto:InstrumentOperationMode* related with information about the operation mode of the instrument; *vsto*:*DateTimeInterval*, being intervals for date and time and *vsto:Parameter*, including the following parameters: *ErrorParameter*, *Noise*, *Period*, *SignalToNoiseRatio*, *TimeDependentParameter* and *StatisticalMeasure*.

The Semantic Web Earth and Environmental Terminology - SWEET Ontology[3] has widespread acceptance in e-Science. However, this ontology extends more in width than depth in certain areas. Thus, for purposes of interoperability and reuse, the *crtdao:MethodType* class was replaced by import of the *sweet:Process* class. It's because the objective of this work is to deepen concepts to generate semantic provenance as the statistical methods used in the analysis of FITS images.

3. Domain Ontologies and Provenance Models

Domain ontologies should be built based on Foundation Ontology, such as SUMO[4], DOLCE[5], UFO[6], among others, because they are theoretically well-founded, becoming the category systems independent of domain, describing the general concepts and improving the quality of conceptual model [Guizzardi 2005]. They are characterized by being highly reusable because it shapes basic and general concepts, as well as relations. However, the well-founded ontologies are generic about many areas.

So, due to the need for representing provenance information, provenance models stands out because are ontologically well-founded representation models, adding concepts and relationships provenance-aware, allowing the adoption of a common provenance terminology [McGuinness et al 2007]. These models are presented follow.

[2] *Virtual Solar-Territorial Observatory:* http://escience.rpi.edu/ontology/vsto/2/0/vsto.owl

[3] *Semantic Web Earth and Environmental Terminology*: http://sweet.jpl.nasa.gov/

[4] *Suggested Upper Merged Ontology*: http://www.ontologyportal.org/

[5] *Descriptive Ontology for Linguistic and Cognitive Engineering* http://www.loa.istc.cnr.it/DOLCE.html

[6] *Unified Foundational Ontology*: http://code.google.com/p/ufo-nemo-project/

3.1 Open Provenance Model - OPM

It is an abstract model developed from Provenance Challenge Series to explain how artifacts were derived, based on workflows. It is independent of technology for interoperability purposes. Uses a graph based on a syntactic rules set and topological constraints. It presents as concepts *Agent*, denoting people; *Process*, denoting actions or executions of process; and *Artifacts*, denoting the entity produced or manipulated. This data model has applicability mainly in biologic area.

The modularity of this data model involves OPM Specification, OPMV Vocabulary, OPMO Ontology and XML Schema. The focus is on provenance in workflows, defining a small set of key concepts to general entities and relationships (*wasGeneratedBy* - WGB and *WasControledBy* - WCB) in workflows. On the downside, the OWL Profile is still evolving to adapt the OPM Specification.

3.2 Provenir

This ontology presents as main concepts *Agent*, *Process* and *Data*. *Data_Collection* and *Parameters* spatial, domain and temporal are subclass of the Data. It is constituted by eight classes and eleven properties, including the Relation Ontology.

It presents as characteristics a common model to represent provenance, being expressive as the concepts and relationships modeled named well-defined, can be extended to modeling of complex provenance information and domain-specific, enabling analysis in SWRL and W3C Rule Interchange Format - RIF. This ontology has applicability in biomedical and oceanography areas in real projects of the e-Science.

3.3 Provenance Markup Language - PML

PML is based on Proof Theory and constitutes a common model for represent and share explanations generated by various intelligent systems such as answers systems of hybrid web questions, analytical text, theorem provers, among others. It describes the justifications as a sequence of information manipulations steps used to generate a response. This sequence is referred as a proof.

Due to modularity, it is possible to use modules individually for *Provenance* (PML-P), *Justification* (PML-J) or hold *Trust* (PML-T) in the data[7]. The PML-T supports annotation of complex trust relations in provenance concepts and justifications. The primitive concepts and relations are specified in OWL, facilitating reuse and extension. The modules PML-P and PML-J are described in the following.

3.3.1 Provenance Ontology - PML-P

PML provides a vocabulary for justification of metadata whose focus is on representational primitives used to describe properties of 'things' identified as information, language and resources, such as organization, person, agent and services. These primitives are extensible, used to annotate the source of information, as to represent sources used and who encoded the information. PML-P presents the following concepts.

[7] URL: http://inference-web.org/2007/primer

An instance of *IdentifiedThing* refers to a real world entity and its properties note the properties of entities such as name, description, date and time of creation and ownership. PML-P also includes *Information*, *Source and SourceUsage*, *Language* and *InferenceRule* subclasses.

The *Information* subclass supports references to information on various levels of granularity and structure, such as a formula in a logical language, a fragment of natural language or a dataset. The *Source* is extensible and refers to a container of information, such as a Document, an Agent, among others. *SourceUsage* is used to associate *Information* and *Source*, declaring information from a *Source* at certain time. *Language* represents the language in that the conclusion is represented. *InferenceRule* aims to encode various types of computation steps.

3.3.2 Justification Ontology - PML-J

This module requires concepts to represent conclusions, zero or more sets of antecedents of the conclusion and the steps used to manipulate information to get conclusions from the set of antecedents and so on recursively. The vocabulary for explanations of data focuses on representational primitives used to explain dependencies between 'things', including constructors to represent how conclusions are derived. It presents the *NodeSet* and *InferenceStep* concepts.

The *NodeSet* represents a conclusion and a set of alternative steps, each of which may provide an alternative justification for a conclusion. This term captures the concept of a set of nodes in steps from one or more proof trees deriving the same conclusion.

An *InferenceStep* represents a justification for the conclusion of the respective *NodeSet*. It refers to a logical step of inference, an information extraction step, any step in the process of computing, or an assertion of a fact or an assumption. It can also be a complex process as web service or application. An *InferenceStep* represents the details such as the *InferenceEngine*, *InferenceRule*, and the set of antecedents *NodeSets* of one justification for the conclusion of the corresponding *NodeSet*.

4. Integration of the Domain Ontology with the Provenance Model PML

In this work, we choose to make use of the PML-P and PML-J modules of PML model, mainly because allows us represent and explain how the conclusions were obtained by informing which the inference engine, the rules and the source of information used, as well as due to modular design.

The integration (Figure2) is done using multiple inheritance of the classes as in Zednik el al (2009), where an individual is defined as a type from the provenance model and at least one type from the domain ontology, e.g. the CoRoT instance is defined as belonging to classes *crtdao:Telescope* and the extension *pmlp:Telescope*, being an subclass of *pmlp:Agent* of the *pmpl:Source* class. So, an instance of *crtdao:Telescope* becomes the source used to justify a conclusion from a *NodeSet*. The same classes in *crtdao* e *pmlp* are treated as equivalent classes.

From a *Question* is created the respective *Query*, which is linked to a *NodeSet*, where is stated a conclusion (*Information*) through the *hasConclusion* property. *NodeSet* may have none, one or more *InferenceSteps* stating which *InferenceRule*, *InferenceEngine* and/or *Source* were used, beyond of a list of antecedents *NodeSets*.

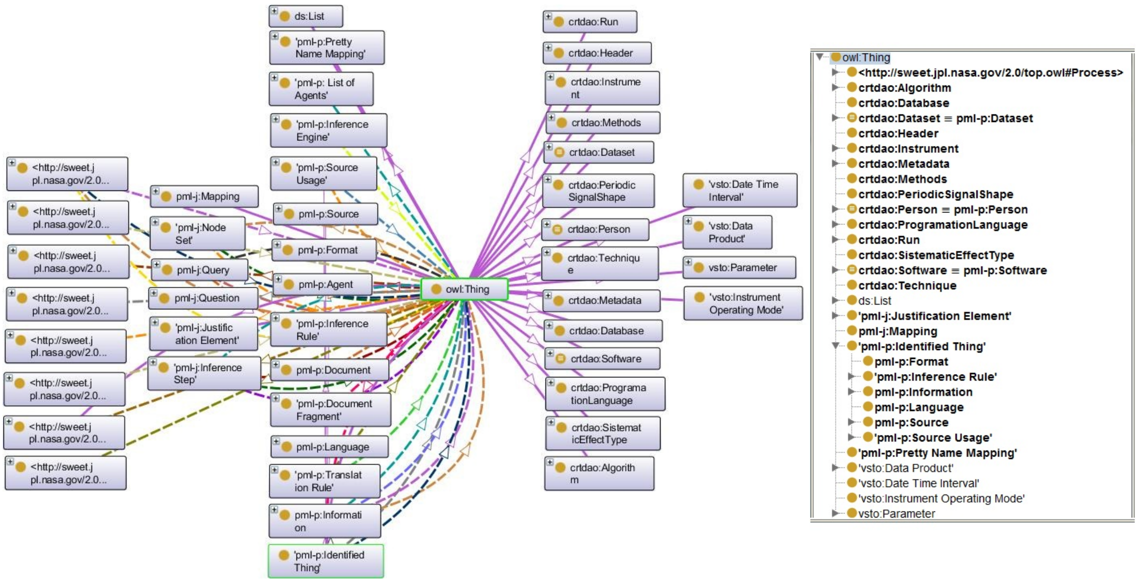


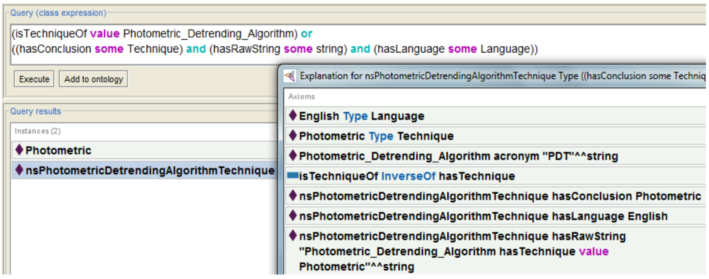
Figure2. Integrated Ontologies visualized in OntoGraf Plugin of Protégé 4.1 Tool

Given the Question '*What is the source of a given dataset*?' is specified the *Query* in a given language binds to a *NodeSet* and declares as conclusion the respective source. As inference step, it is possible to declare the *Source* of the using the property *hasSource*. So, it is possible to justify the source of information in the standardized way.

McGuinness et al (2007) identifies four types of justifications for a given conclusion, exemplified below in XML format and Protégé 4.1 Tool:

i. *The conclusion is an unproven conclusion or goal*. No justification is available and none InferenceStep is associated with the NodeSet. For the Question *What is the technique of the Photometric Detrend Algorithm?* and the Query in the Manchester OWL DL Query syntax (Figure3), the conclusion is given by NodeSet respective using the properties *pmlp:hasLanguage* related with the language of the conclusion and *pmlp:hasRawString* related with the content of information as a string.

```
<pmlj:NodeSet
nsPhotometricDetrendingAlgorithmTechnique>
<pmlj:hasConclusion>
<pmlp:Information>
<pmlp:hasLanguage>(English)>
<pmlp:hasRawString datatype="string">
(Photometric_Detrending_Algorithm
hasTechnique value Photometric)
</pmlp:hasRawString>
</pmlp:Information>
</pmlj:hasConclusion>
</pmlj:NodeSet>
```



(isTechniqueOf value Photometric_Detrending_Algorithm) or
((hasConclusion some Technique) and (hasRawString some string) and (hasLanguage some Language))

Photometric
nsPhotometricDetrendingAlgorithmTechnique

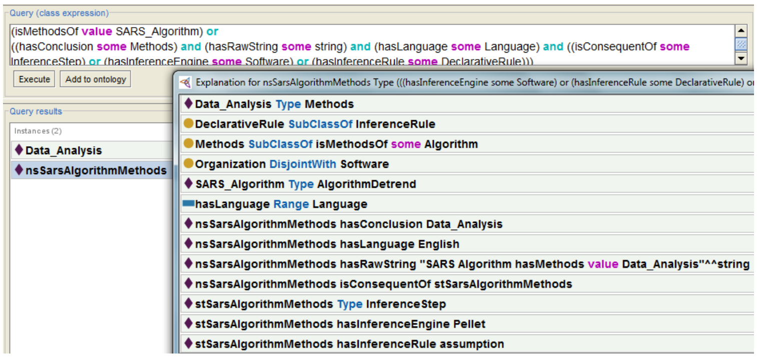
Explanation for nsPhotometricDetrendingAlgorithmTechnique Type ((hasConclusion some Techniq

English Type Language
Photometric Type Technique
Photometric_Detrending_Algorithm acronym "PDT"^^string
isTechniqueOf InverseOf hasTechnique
nsPhotometricDetrendingAlgorithmTechnique hasConclusion Photometric
nsPhotometricDetrendingAlgorithmTechnique hasLanguage English
nsPhotometricDetrendingAlgorithmTechnique hasRawString "Photometric_Detrending_Algorithm hasTechnique value Photometric"^^string

Figure3. Justifying a unproven Conclusion without InferenceStep

ii. *The conclusion is an assumption*. The conclusion is directly assumed by an agent as a true statement. The Question *What Methods the SARS algorithm belongs to?* is justified by inference in the NodeSet respective that includes the information *pmlp:hasRawString* and *pmlp:hasLanguage*. As a consequence of the InferenceStep is declared *assumption* as InferenceRule and *Pellet* as InferenceEngine (Figure4).

```
<pmlj:NodeSet nsSarsAlgorithmMethods>
<pmlj:hasConclusion>
<pmlp:Information>
<pmlp:hasRawString datatype=“string”>
(SARS_Algorithm hasMethods value
Data_Analysis)
</pmlp:hasRawString>
<pmlp:hasLanguage>(English)>
</pmlp:Information>
</pmlj:hasConclusion>
<pmlj:isConsequentOf>
<pmlj:InferenceStep>
<pmlj:hasInferenceEngine>Pellet
<pmlj:hasInferenceRule>assumption
</pmlj:InferenceStep>
</pmlj:isConsequentOf>
</pmlj:NodeSet>
```



```
(isMethodsOf value SARS_Algorithm) or
((hasConclusion some Methods) and (hasRawString some string) and (hasLanguage some Language) and ((isConsequentOf some
InferenceStep) or (hasInferenceEngine some Software) or (hasInferenceRule some DeclarativeRule)))
```

Data_Analysis
nsSarsAlgorithmMethods

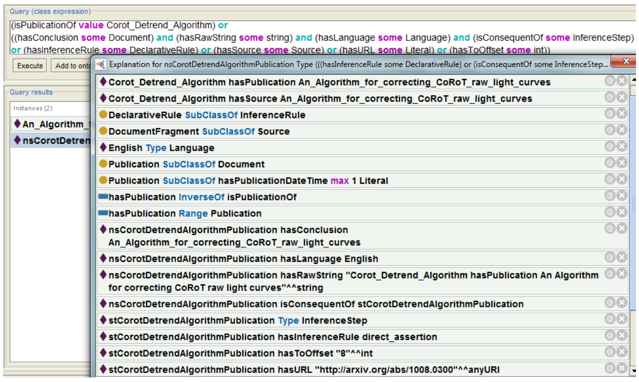
Explanation for nsSarsAlgorithmMethods Type (((hasInferenceEngine some Software) or (hasInferenceRule some DeclarativeRule) or

```
Data_Analysis Type Methods
DeclarativeRule SubClassOf InferenceRule
Methods SubClassOf isMethodsOf some Algorithm
Organization DisjointWith Software
SARS_Algorithm Type AlgorithmDetrend
hasLanguage Range Language
nsSarsAlgorithmMethods hasConclusion Data_Analysis
nsSarsAlgorithmMethods hasLanguage English
nsSarsAlgorithmMethods hasRawString "SARS Algorithm hasMethods value Data_Analysis"^^string
nsSarsAlgorithmMethods isConsequentOf stSarsAlgorithmMethods
stSarsAlgorithmMethods Type InferenceStep
stSarsAlgorithmMethods hasInferenceEngine Pellet
stSarsAlgorithmMethods hasInferenceRule assumption
```

Figure4. Justifying a Conclusion using InferenceStep

iii. *The conclusion is a direct assertion*. It can be declared by Inference Engine directly without using any antecedent information (Figure5). For the Question *What is the publication of Corot Detrend Algorithm?*, the *NodeSet* respective declare the information using *pmlp:hasRawString* and *pmlp:hasLanguage* properties. As the consequence, the *InferenceStep* informs *direct_assertion* using *pmlj:hasInferenceRule* and the *pmlp:hasDocument* property informs the *Source*. Also it is possible to declare other details about the publication how number of pages and URL.

```
<pmlj:NodeSet nsCorotDetrendAlgorithmPublication>
<pmlj:hasConclusion>
<pmlp:Information>
<pmlp:hasRawString datatype=”string”>
(Corot Detrend Algorithm hasPublication value An algorithm for
correction CoRoT raw light curves)
</pmlp:hasRawString>
<pmlp:hasLanguage>(English)>
</pmlp:Information>
</pmlj:hasConclusion>
<pmlj:isConsequentOf>
<pmlj:InferenceStep>
<pmlj:hasInferenceRule>direct_assertion
<pmlp:SourceUsage>
<pmlp:hasDocument>An algorithm for correction…
<pmlp:hasFromOffset>1</pmlp:hasFromOffset>
<pmlp:hasToOffset>8</pmlp:hasFromOffset>
…
```



```
(isPublicationOf value Corot_Detrend_Algorithm) or
((hasConclusion some Document) and (hasRawString some string) and (hasLanguage some Language) and (isConsequentOf some InferenceStep)
or (hasInferenceRule some DeclarativeRule) or (hasSource some Source) or (hasURL some Literal) or (hasToOffset some int))
```

An_Algorithm_
nsCorotDetren

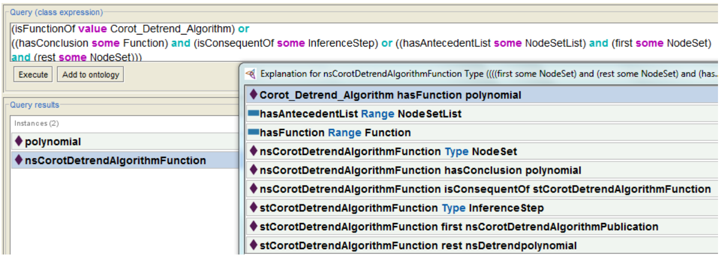
Explanation for nsCorotDetrendAlgorithmPublication Type (((hasInferenceRule some DeclarativeRule) or (isConsequentOf some InferenceStep...

```
Corot_Detrend_Algorithm hasPublication An_Algorithm_for_correcting_CoRoT_raw_light_curves
Corot_Detrend_Algorithm hasSource An_Algorithm_for_correcting_CoRoT_raw_light_curves
DeclarativeRule SubClassOf InferenceRule
DocumentFragment SubClassOf Source
English Type Language
Publication SubClassOf Document
Publication SubClassOf hasPublicationDateTime max 1 Literal
hasPublication InverseOf isPublicationOf
hasPublication Range Publication
nsCorotDetrendAlgorithmPublication hasConclusion
An_Algorithm_for_correcting_CoRoT_raw_light_curves
nsCorotDetrendAlgorithmPublication hasLanguage English
nsCorotDetrendAlgorithmPublication hasRawString "Corot_Detrend_Algorithm hasPublication An Algorithm
for correcting CoRoT raw light curves"^^string
nsCorotDetrendAlgorithmPublication isConsequentOf stCorotDetrendAlgorithmPublication
stCorotDetrendAlgorithmPublication Type InferenceStep
stCorotDetrendAlgorithmPublication hasInferenceRule direct_assertion
stCorotDetrendAlgorithmPublication hasToOffset "8"^^int
stCorotDetrendAlgorithmPublication hasURL "http://arxiv.org/abs/1008.0300"^^anyURI
```

Figure5. Justifying a Conclusion using Direct Assertion

iv. *The conclusion is derived from a list of antecedents by applying a certain computation*. This representation to encode many types of computation steps. The Question *Which is the function type of the Corot Detrend Algorithm?* (Figure6) shows that the conclusion is derived from first NodeSet or from rest NodeSet.

```
<pmlj:NodeSet nsCorotDetrendAlgorithmFunction>
<pmlj:hasConclusion>
<pmlp:Information>
<pmlp:hasLanguage>(English)>
<pmlp:hasRawString>Corot_Detrend_Algorithm
hasFunction polynomial
</pmlp:Information>
</pmlj:hasConclusion>
<pmlj:isConsequentOf>
<pmlj:InferenceStep>
<pmlj:hasAntecedentList>
<pmlj:NodeSetList>
<ds:first>nsCorotDetrendAlgoritmPublication
<pmlj:hasIndex>0</hasIndex>
</ds:rest>nsDetrendpolynomial
<pmlj:hasIndex>1</hasIndex>
…
```



```
(isFunctionOf value Corot_Detrend_Algorithm) or
((hasConclusion some Function) and (isConsequentOf some InferenceStep) or ((hasAntecedentList some NodeSetList) and (first some NodeSet)
and (rest some NodeSet)))
```

polynomial
nsCorotDetrendAlgorithmFunction

Explanation for nsCorotDetrendAlgorithmFunction Type ((((first some NodeSet) and (rest some NodeSet) and (has...

```
Corot_Detrend_Algorithm hasFunction polynomial
hasAntecedentList Range NodeSetList
hasFunction Range Function
nsCorotDetrendAlgorithmFunction Type NodeSet
nsCorotDetrendAlgorithmFunction hasConclusion polynomial
nsCorotDetrendAlgorithmFunction isConsequentOf stCorotDetrendAlgorithmFunction
stCorotDetrendAlgorithmFunction Type InferenceStep
stCorotDetrendAlgorithmFunction first nsCorotDetrendAlgorithmPublication
stCorotDetrendAlgorithmFunction rest nsDetrendpolynomial
```

Figure6. Justifying a Conclusion derived from AntecedentList

5. Related work

Zednik et al (2009) present how the semantic provenance is reconstructed to data products in coronal physics area. This work provides a foundation for scientific workflow provenance applications, describing the use of semantic web technologies to encode provenance and domain information and demonstrating how both can be used together to satisfy complex use case. The data model use OWL ontologies independent. The Solar-Terrestrial Ontology – VSTO is used as a core domain model in e-Science, modeling data products, instruments and parameters. The provenance model uses the Inference Web and the Framework PML, chosen because of its capabilities of represent conclusions, justificatives and explanations. The integration of provenance and domain models is done by means of multiple-inheritance from individuals´ declarations of the ontologies. The search results can be seen by Inference Web browser or by Probe-It!, enabling scientists to better understand imperfections and processing consequences upon e-Science data images.

Malaverri et al (2012) presents an approach of provenance to ensure the quality of geospatial data, combining features provided by the OPM and FGDC geographic metadata standards. It presents a case study in agriculture area, considering the trustworthiness of source, is that, the degree of confidence of who created/made available the data and temporality dimensions including valid and transaction time, e.g. 'when' related to data quality. Despite the proposal model be based on OPM model, is added own characteristics taking into account the geospatial domain and assessment of data quality. As future works, techniques to compute and assess the trustworthiness of data will be investigated.

Salayandia et al (2012) propose a framework to support the creation of ontologies for management of scientific data, specifying an abstraction in the form of a top-level ontology codified in OWL-DL, including general concepts that can be specialized to describe the capture and transformation of data. The Ontology Driven Workflow (WDO.owl) is proposed and presents three basic concepts: Date, things that can be used directly or indirectly as evidence, e.g. the output of a sensor; Method, things that can be used to transform the data, e.g. visualization of software; and Container, things that can be used as acquires or placeholders of the data, e.g. a database. WDO is specified in Description Logic and the knowledge representation system is divided into Tbox terminology and Abox, including assertions as the individuals in relation to the Tbox. WDO is aligned with PML, where the concepts Date, Container and Method are included respectively by PML concepts: Information, Source and Inference Rule. The formalism that aligns the WDO and PML Ontologies is also specified using DL, including subsumption equations rather than equalities due to concepts related with the provenance are more general than the concepts of the WDO Ontology. It is because data can be transformed by systematic processes, where the framework can be used to document the process.

This paper stands out by enrich with semantic and standardization the phases of the detrending and exoplanets search, providing information about the semantic provenance of data and statistical methods used in the correction and analysis of FITS images, contributing for adding semantic knowledge in experiments of e-Science and take advantage of the features provided by PML.

6. Conclusions

It is presented in this paper that is possible to generate semantic provenance in scientific images analysis. The environment involves FITS images from CoRoT Archive and the integration of data models related to domain ontology and the provenance model. This integration allow us make use of a common model and standardized for generating provenance, contributing for semantic interoperability and allowing us to justify how conclusions were obtained in the knowledge base.

Due to need for representing provenance information, provenance models are ontologically well-founded, adding concepts and relationships provenance-aware, allowing the adoption of a common provenance terminology. In this work, we choose to use the PML model by allowing us to represent and explain how the conclusions were obtained providing the inference engine, the inference rules and the source used, as well as due to its modularity.

The semantic provenance information obtained will be persisted in databases, and integrated in a web framework, facilitating the information retrieval processes, where queries of provenance can be performed, allowing further analysis and contributing to enrich semantically the development of scientific experiments. Despite the scope of this work, results can be expanded to fields of e-Science where the scientific images analysis requires preprocessing, adding semantic knowledge and allowing interoperability.

Acknowledgments

We acknowledge the support of the Astronomical Observatory AstroUEPG.

References

Baader, F. and Calvanese, D. and McGuinness, D. L. and Nardi, D. and Patel-Schneider, P. F. (2003). The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, New York, USA.

Borst, W. N. (1997). Construction of Engineering Ontologies for Knowledge Sharing and Reuse. Doctoral Thesis. University of Tweenty, Centre for Telematica and Information Technology, Enschede, The Netherlands, (227 pp.).

Buneman, P. and Khanna, S. and Tan, W. C. (2007). Why and Where: A Characterization of Data Provenance. In ICDT (pp. 316–330).

Cheney, J. and Chiticariu, L. and Tan, W. C. (2009). Provenance in Databases: Why, How, and Where. Foundations and Trends in Databases (Vol. 1, N. 4, pp. 379–474). Hanover, MA, USA: Now Publishers Inc.

Davidson, S. B. and Freire, J. (2008). Provenance and Scientific Workflows: Challenges and Opportunities. In Proceedings of ACM SIGMOD (pp. 1345–1350).

De Souza, L. and Vaz, M. S. M. G. and Emílio, M. and da Rocha, J. C. F. and Boufleur, R. (2011). Data Analysis Provenance: Use Case for Exoplanet Search in CoRoT Database. Presented In: ADASS XXI Conference Series. In Press: ASP Conf. Ser. Vol. TBD. San Francisco. 2012.

Guizzardi, G. (2005) Ontological Foundations for Structural Conceptual Models. PhD Thesis (CUM LAUDE), University of Twente, The Netherlands. Published as the book Ontological Foundations for Structural Conceptual Models, Telematica Instituut Fundamental Research Series No. 15.

Hanisch, R. J. and Farris, A. and Greisen, E. W. and Pence, W. D. and Schlesinger, B. M. and Teuben, P. J. and Thompson, R. W. and Warnock III, A. (2001). Definition of the Flexible Image Transport System (FITS). A&A (Vol. 376, pp. 359–380).

Horrocks, I. and Kutz, O. and Sattler, U. (2006). The Even More Irresistible SROIQ. Proc. of the 10th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2006), (pp. 57–67), AAAI Press.

Horrocks, I. and Patel-Schneider, P. F. and Bechhofer, S. (2005) OWL Rules: A Proposal and Prototype Implementation. Journal of Web Semantics, V. 3, N. 1.

Knublauch, H. and Fergerson, R. W. and Noy, N. F. and Musen, M. A. (2004). The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. 3rd ISWC 2004, Hiroshima, Japan. (pp. 229–243).

Malaverri, J. E. G. and Medeiros, C. B. and Lamparelli, R. C. (2012) A provenance Approach to Assess Quality of Geoespatial Data. Symposium on Applied Computing.

McGuinness, D. L. and Ding, L. and da Silva, P. P. and Chang, C. (2007). PML 2: A Modular Explanation Interlingua. In Proc. of the AAAI 2007 Workshop on Explanation-aware Computing (pp. 22–23).

Moreau, L. and Clifford, B. and Freire, J. and Futrelle, J. and Gil, Y. and Groth, P. and Kwasnikowska, N. and Miles, S. and Missier, P. and Myers, J. and Plale, B. and Simmhan, Y. and Stephan, E. and den Bussche, J. V. (2011). The Open Provenance Model Core Specification (v1.1). Future Generation Computer Systems (Vol. 27, N. 6, pp. 743–756). Amsterdam, The Netherlands: Elsevier Science Publishers B. V.

Noy, N. F. and McGuinness, D. L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Tech. Rep. KSL-01-05 and Stanford Medical Informatics Tech. Rep. SMI-2001-0880 (25 pp.).

Sahoo, S. S. and Sheth, A. and Henson, C. (2008). Semantic Provenance for eScience. IEEE Computer Society, pp. 46-54.

Sahoo, S. S. and Weatherly, D. B. and Mutharaju, R. and Anantharam, P. and Sheth, A. and Tarleton, R. L. (2009). Ontology-Driven Provenance Management in eScience: An Application in Parasite Research. Proc. of the CoopIS, DOA, IS, and ODBASE 2009 on On the Move to Meaningful Internet Systems: Part II. OTM ’09. (pp. 992–1009). Berlin, Heidelberg: Springer-Verlag.

Salayandia, L. and Pinheiro, P. and Gates, A. Q. (2012). A Framework to Create Ontologies for Scientific Data Management. University of Texas at El Paso.

Tan, W. C. (2007). Provenance in Databases: Past, Current, and Future. IEEE Data Eng. Bull. (Vol. 30, N. 4, pp. 3-12).

Zednik, S. and Fox, P. and McGuinness, D. L. and da Silva, P. P. and Chang, C. (2009). Semantic Provenance for Science Data Products: Application to Image Data Processing. First International Workshop on the role of Semantic Web in Provenance Management.