Semantic interoperability in astrophysics for workflows extraction from heterogeneous services

Thierry Louge, Mohamed Hedi Karray, Bernard Archimède, Jürgen Knödlesder

Semantic interoperability in astrophysics for workflows extraction from heterogeneous services

Thierry Louge[1], Mohamed Hedi Karray[2], Bernard Archimède[2], Jürgen Knödlseder[3]

[1] Université de Toulouse - UPS, Institut de Recherche en Astrophysique et Planétologie, 57 avenue d'Azereix, 65000 Tarbes, France
[2] Université de Toulouse, Ecole Nationale d'Ingénieurs de Tarbes ENIT, 47 avenue d'Azereix, BP1629, 65016 Tarbes Cedex, France
[3] CNRS, Institut de Recherche en Astrophysique et Planétologie, 9 avenue du colonel Roche, BP 44346, 31028 Toulouse Cedex 4, France

Abstract. Modern instruments in astrophysics lead to a growing amount of data and more and more specific observations, among which scientists must be able to identify and retrieve useful information for their own specific research. The Virtual Observatory (VO)[1] architecture has been designed to achieve this goal. It allows the joint use of data taken from different instruments. Retrieving and cross-matching those data is in progress, but it's impossible today to find a sequence resolving a given science case needing a combination of existing services of whom the user doesn't knows the specifications. The goal of this work is to propose the basis of an architecture leading to automatic composition of workflows that implement scientific use cases.

Key words: interoperability, ontologies, virtual observatory, astrophysics

1 Introduction

In view of the ever-growing quantity of scientific data provided by modern astrophysics, the community of universe sciences built a system of "virtual" observatories, allowing to express metadata in a shared format (VOTable[2] being the most widely used) and offering a set of protocols and services to access the data. The goal of the associated architecture is allowing the share of scientific data produced by instruments from all fields of universe sciences, from astrophysics to geophysics through planetology, heliophysics, etc. The global goal is very well shared by everyone involved but many specific needs occured, sometimes leading to specific developments ending with the emergence of several VO "branches", guided by different organisations such as IVOA (International Virtual Observatory Alliance) for astrophysics, VAMDC (Virtual Atomic and Molecular Data

[1] http://www.ivoa.net/deployers/intro_to_vo_concepts.html
[2] http://www.ivoa.net/documents/VOTable/

Center) for astrochemistry, IPDA (International Planetary Data Alliance) for planetology, etc. Furthermore, the volume of data increases in every science field and the needs for common protocols and formats are shared outside of astrophysics. In this context, Research Data Alliance[3] deals with the same kind of challenges than the VO, in order to organize every science field around the same concepts and sofware architecture. Expressing data and services in a shared format should lead to an easier way to find and combine appropriate services for scientific uses.

In the field of services computing research, a common way used to find web services is to use Service Oriented Access Protocol (SOAP), in cunjunction with Web Services Definiton Language (WSDL) services descriptions and Universal Description Discovery and Integration (UDDI) registries[15] to locate appropriate services. This approach is expected to reach a new level of effectiveness with the emergence of semantic web principles [14], and the use of ontologies describing knowledge under the form of metadata with concepts, relationships and objects.

We present in this work an architecture combining the methods used on service discovery and contributions of the VO in astrophysics. This architecture allows a VO transparency enhancement by performing the matching and selection of services automatically, from the description of a scientific use-case. We should be able to combine in our workflows VO and non VO-related services alike, providing that they are correctly described in the ontology and detected as relevant for the given use-case. Generated workflows will be presented to the user who'll be able to closely inspect every single step to evaluate the results, judge the accuracy and annotate them to provide enhancements for future or immediate re-runs. In this paper, we'll briefly expose the state of the art concerning web services composition and VO capabilities, then suggest an architecture to allow automation of workflows composition and the first test results we get.

2 State of the art

2.1 Web services composition

A way to resolve web services composition is to query a UDDI services registry, select appropriate services based on their WSDL description and query them with SOAP protocol. "WSDL is the emerging language for describing the present web service technology and presents the syntactic description of the web services. It only present the structure of the data sent and received through the web, but is unable to present the meaning of the data" [17]. Such a description, focusing on the semantics of data rather than their technical representation may be obtained using ontologies.

Ontologies may be used as interoperability layer between services, to ensure that skills of one service corresponds to the needs of another one[16]. More

[3] https://rd-alliance.org

specifically, ontologies are used to describe services, the way they operate and the data they need to be used. One of the purposes is to increase the effectiveness of interoperablity, selection and composition of services by describing them in one common ontology, which is very close to what we would like to realize with astrophysics services and that we present in section 4.

Semantic web makes software agents regular web users as are humans, and enhance web services composition thanks to the new reasoning possibilities offered, as exposed in [18]. In this paper, authors expose several existing approaches dealing with services composition and conclude that inputs and outputs of services are not enough to get an appropriate composition. In order to enhance composition performance, one has to specify the services pre/post conditions. The pre-condition prescribes what is necessary to hold before the Web service can be executed and the post-condition prescribes what holds after the service execution[18]. This combination of compatibilities, pre-execution conditions to match and post-execution results to achieve is completed with the notion of Quality of Service (QoS) describing how non-functional requirements have been judged during the execution of the service (response time, availability...).

Then authors review several approaches for web composition, like using Knowledge Interchange Format (KIF) rules to express user constraints to match with an ontology for services (OWL-S), which is the closest one to the architecture that we present in this work.

2.2 Virtual Observatory (VO) in astrophysics

VO is a software construction very tied to its application domain that allows to express observed and theoretical data with a common description, and the building of services based on the same formats and protocols. Interoperability, which is the core concern of this architecture is reached through definite descriptive fields and software tools able to understand the VO formats, datamodels and protocols. Nevertheless, difficulties still exist because of the multiple different ways to adapt the datamodels, imposed by the great amount of specific definitions tied to specific observations and their diversity. Even if the VO is nowadays a reality and a success, its everyday use is frequently restricted by not providing enough ease-of-use, because of a too weak transparency for the end user that has to deal with thousands of services with little support or poor descriptions.

Datamodels: the description. Querying a VO service returns an XML document, which is called "datamodel" (DM), and defines the mandatory information so that the answers of a service can be used by VO-compliant software, and optional information completing the minimal required description. Datamodels can be used by different protocols and share some vocabulary to interoperate.

Software querying the VO must, to be able to properly use the data, undestand every DM.

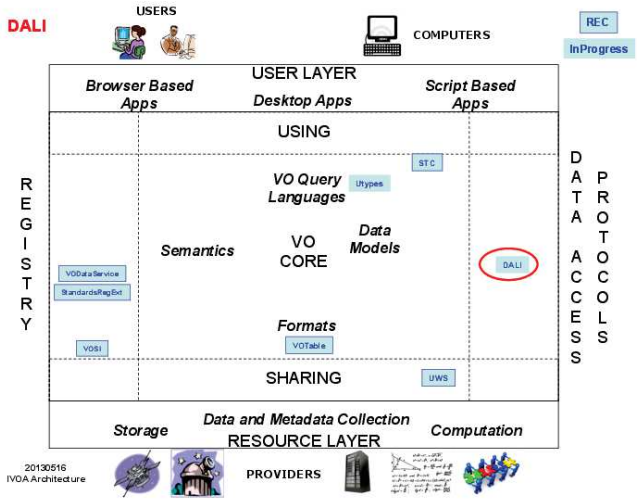

Fig. 1. Data Access layer (DAL) in the overall IVOA architecture

Protocols: data access. As shown in figure1[4], IVOA data access layer is composed of several protocols, each of them being dedicated to a service category such as Simple Spectral Access (SSA) for spectra, Table Access Protocol (TAP) for catalogs of observations or direct access to database tables, etc. Generally, protocols are not tied to specific DMs, with the exception of SSA which relates to the spectra DM.

An example is the ConeSearch protocol, which is rather widely used and implemented by a large number of services, and which allows to search for an observation in the very general term, being a spectra, an image or anything else, real or theoretical around a reference point in the sky. As ConeSearch allows to describe data in a very general way it allows to retrieve any kind of observation and so any kind of scientific results. Today, more than ten thousand different services are registered serving this ConeSearch protocol, and the diversity of their results and specificities is a burden for an effective interoperability.

Semantic interoperability in IVOA: UTypes , UCDs, VOUnits. Data description in the DM use a defined ensemble of symbols (UTypes) referencing information that can be found inside the structure of the given DM, coupled with a more generic vocabulary allowing the user to get some details about the given information: the UCDs (Universal Content Descriptors).

IVOA data description is completed by another recommandation (VOUnits), listing every unit understandable by VO-compliant tools, and suggests to simply put non-listed units between single-quotes.

[4] http://www.ivoa.net/documents/DALI/20131129/REC-DALI-1.0-20131129.html

This can be illustrated with an example coming from Photometry DM: we find UTYPe "photDM: PhotometryFilter.spectralLocation.unit.expression" designing "Unit of the spectral axis used to characterize the spectral ccordinate of the zero point" associated with the ucd "meta.unit" designing the unit. In an SSA answer from a service we could find: "ucd="instr.bandwidth" utype="SSA:Char.SpectralAxis.Coverage.Bounds.Extent "unit="angstrom"", for the meaning of the information (ucd), its role in the DM (utype) and its unit (unit). Despite all those possibilities, some specific data are not taken into account by the DM definitions, hence some information is lost as there is no equivalent VO representation, and the corresponding services can not used in an interoperable way. An example are polarized spectra: while spectra can be described using the spectrum DM, there exists no description for the polarization information, neither at the DM level or the service description, which stronly limits the usage of the data.

Also, we frequently find services with only partial use, or non-standard use of the DMs (one frequent case is to meet ucd="POS_EQ_RA" for pos.eq.ra which is the correct ucd) as the data provided are not systematically checked.

All these reasons call for the addition of an interoperability layer, as implemented for example in the IRIS framework[10], allowing to attach supplementary information to VO services.

Software tools Dedicated software[5] exists allowing the query of VO registries and retrieval and understanding of data. Sometimes very general as Aladin, or more specialized ("Montage" for images mosac vizualisation, "CASSIS" for the vizualisation of spectra, just to cite those ones), they are the interface between users and the mechanisms described above. Sometimes, they only serve a predefined ensemble of services[6], for which their performances are optimized and the precise data description known beyond the DM content. Software development, specific to a certain kind of data categories are regurlarly appearing, such as photometry in the Vizier catalogs[1].

Another kind of tools that exist are the workflows planners. They offer an automatization of workflows composed of queries to predefined VO services and scientific processing. The principle is that the user defines a solution to the problem, builds a workflow by specifying what services are to query and how data are to be processed with which tools. The workflows can be executed as often as required, for example with different input parameters, and they can be publicly shared with the scientific community (e.g. http://www.myexperiment.org). Taverna is one of those tools and integrated in some of the HELIO (heliophysics-oriented VO) services to provide the user direct description of HELIO services inside Taverna quickly and easily[2].

These considerations on data discovery were met again concerning the scientific software and lead to the idea of having an application registry that would allow to access directly the tools that fit the user needs. Initiatives such as Astrophysics Source Code Library (ASCL), which development is still on progress[13]

[5] http://www.ivoa.net/astronomers/applications.html
[6] http://www.usvao.org/science-tools-services/time-series-search-tool/

aims at providing such a registry. One of the main difficulties for the users today is indeed to locate and learn to use the appropriate tool for a scientific use-case, and to put it in relation with other software tools if needed.

3 Practical use of the VO.

3.1 Using the VO: Overview.

The data models used by IVOA are both flexible and heterogeneous. Mandatory keywords are limited, but necessarily imprecise to allow adaption to a large variety of data or different origins. Each service can enrich the description according to the defined format, yet there is no guarantee that all services will implement the same extensions. For some areas of research (e.g., gamma-ray astronomy), the possibilities for describing observations are limited. Therefore initiatives as HELIO[2] to heliophysics appear, trying to provide a more accurate description of specific data. Another problem is the knowledge of the existence of services. Current registries provide a list of services and their characteristics, but this list may be very long, making it difficult for a user to identify the most adapted service for a given use-case.

Even in the case of two services offering the same type of data (spectra, for example) and in the same wavelength band, there is nothing to put both in relation, and a user accessing one of the services will not be informed about the existence of the second. These concerns are taken into account by the IVOA, which works on the development of a protocol called "DataLink" [7]. Once established, DataLink will allow a data provider to specify other data in relation to those it provides. However, this link will be established based on knowledge of a data provider and according to the capacity of each organization to provide this protocol, to maintain and update its content from the emergence of new data and / or new services.

So it is the user's responsibility to make a selection and ensure the joint use of data, which can be a complex operation due to the large amounts of data and data sources that exist. This large number of possibilities involves treatment "a priori" by the user, which lead primarily to already known services, and can not sort of the more than 10,000 service offering for example the protocol "Cone Search" what are those likely to provide useful information to its study. The concrete and systematic use of the Virtual Observatory remains complicated for including an informed user due to differences between the technical sales descriptions of services and their multiplicity.

3.2 Use-case: Analysis of the Crab nebula.

Let us consider a specific use-case for reference: an astrophysicist wants to produce a multi-wavelength analysis of the Crab Nebula. This case study is similar

http://www.ivoa.net/documents/DataLink/20140930/PR-DataLink-1.0-20140930.html

to a case described in an article in the SF2A (French Society of Astronomy and Astrophysics) [3], which searches for the same type of analysis on two services, HESS and Fermi-LAT.

How to get there with current software? The first step consists of using a tool that queries OV services to identify those that provide spectra. Spectral data can be provided by services satisfying the ConeSearch and the SSA protocols. Both protocols need to be examined. For services satisfying the ConeSearch protocol those have to be identified which according to the provided UCDs actually provide spectral information. From the resulting list of services, a detailed analysis of the service description needs to be made to identify the services that are relevant to the problem (e.g. which are the data of highest quality, which data are obsolete, which data are inaccurate, etc.). Also, the services need to be identified that provide data in a format and in units that are exploitable by the tools at hand. Doing so on hundreds or thousands of services is impossible without automation, and we likely will choose the first we meet and seem to agree about. Then, the user can recover the data provided that the server is not down, and provided that the actual source of interest (here the Crab Nebula) has actually been observed. Eventually at this step, alternative services need to be considered.

4 Proposed solution

4.1 Design of an astrophysics services ontology.

As we have seen, the VO covers multiple aspects and although we took the IVOA as an example of architecture, yet not all astrophysical information and services do comply to VO standards. Our goal is to develop a solution that uses the Virtual Observatory as transparent as possible so that an end user would not be concerned about data query and reading, service identification, and mixing VO and non-VO services. In the world of bioinformatics, a similar problem of interoperability is addressed by the SADI project [6], a web service description model based on Ontology Web Langage (OWL) for particular services to interface with Taverna. Our approach has many similarities with this work, extending its principle to the workflows OWL description and to place the OWL representation services outside of the services themselves, to allow existing models to continue to operate without changes and to integrate into the system.

The overall architecture of our system is illustrated in Figure 2. The ontology that we will produce and will be updated by different sources, OVs and autonomous services alike. It will generate a knowledge base within which the reasoning will take place. The results of our work are intended to be used as a web service to various input levels:

• The standard user, who will provide scientific cases for which we propose processing streams available.

• Second level of the user wish to consult the knowledge at his disposal by visualizing ontology and performing queries on its structure.

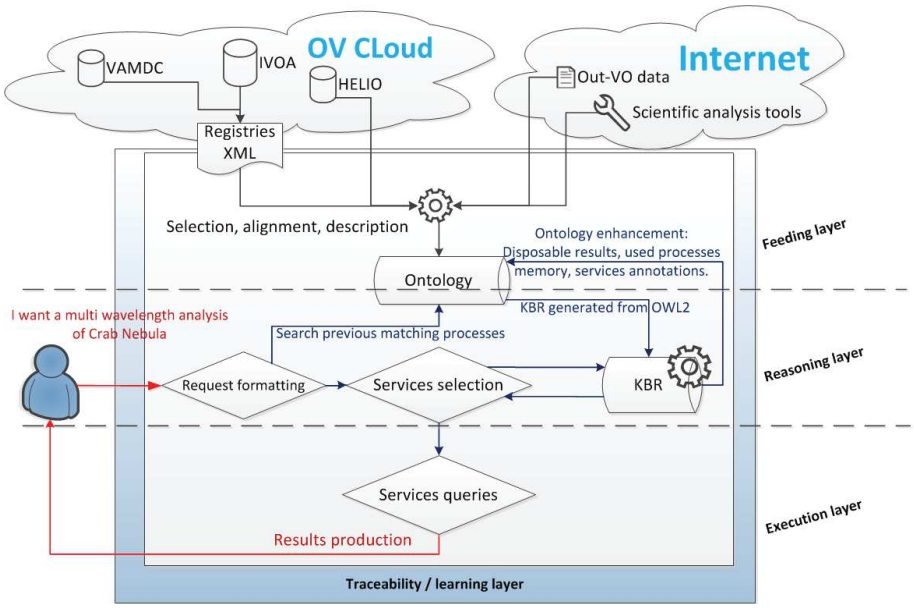

Fig. 2. Overall architecture

• The third level of user, who can enter the description of a service in the ontology, to see it incorporated into the range of opportunities available.

• The administrator, who will update the ontology with new treatment libraries or tools at large installed on the physical server and use program, and descriptions from the third level user alike.

The ontology will be used through a web interface, and updated by the administrator of the system based on suggestions from the users concerning service or workflow comments and annotations, or new services candidates to be part of the system.

4.2 Structure and ontology filling.

Figure 3 focuses on the main source of knowledge in the ontology, which are the description of the skills of web services. They are either collected through XML descriptions issued from registries (IVOA organization) by the module "ASTRO1" or through other available documents (WSDL-like descriptions, and the system will also provide a specific interface dedicated to descriptions of new services). After being collected, the description is analyzed to gather informations concerning the skills of the service and to detect whereas and information (provided or needed by the service) is already known in the system or is a new one (module ISC, Individuals Selection and Comparison). Finally, the service is put on OWL2 description and integrated into the ontology (module DEUS, DEscribe and Update Services).

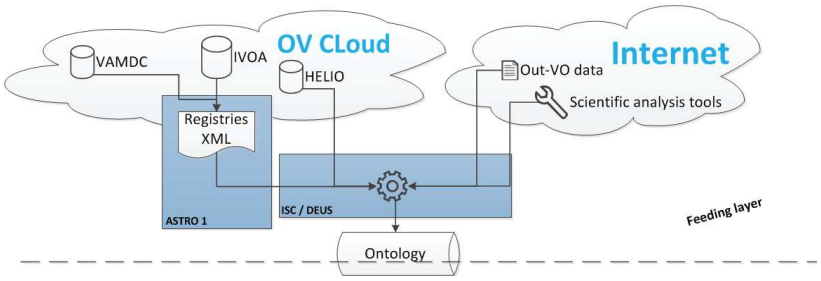

Fig. 3. Feeding layer modules

The structure of knowledge represented in the ontology must be free of technical elements, even if it must be able to ensure the orchestration of elected treatment waves, to go to programs that can query the various protocols (in the case of services derived from OV) and query interfaces of autonomous web services. The structure of the ontology used to represent domain knowledge, support for both the description contained in the existing data models and skills of available services. The workflows generated by the system will also be included in this structure. Services and workflows are described with the same semantic metadata, which support interoperability between the collected data.

Matching the informations. A more detailed description on the ASTRO1, ISC and DEUS modules is shown in Figure 4. When an existing service provides new or updated information, or when a new service becomes available the ontology needs to be updated. This implies matching the new information with any existing information to identify to what class the new information belongs, or if an updated information needs to be merged with some already existing information.

This identification of ontology elements to link with new sources of information is an important aspect for the sustainability and genericity of our system. Our design will rely on the principles of finding alignments between concepts based on their descriptions and mapping semantic models, learning from them to best understand furthers ones that have been outlined in references[7, 8].

4.3 Reasoning with the ontology.

Request representation. The questioning of this service go through the interpretation of the requests made by the user to understand the elements of the system. The reconciliation between the expression of the case by the user and the concepts and relations of the ontology will be managed by assistance to the collection and use of key words recognition techniques based on parsing the natural langage [11]. In addition with the natural expression, an interface-driven query construction will help the user to describe the use-case he wants the system to solve.

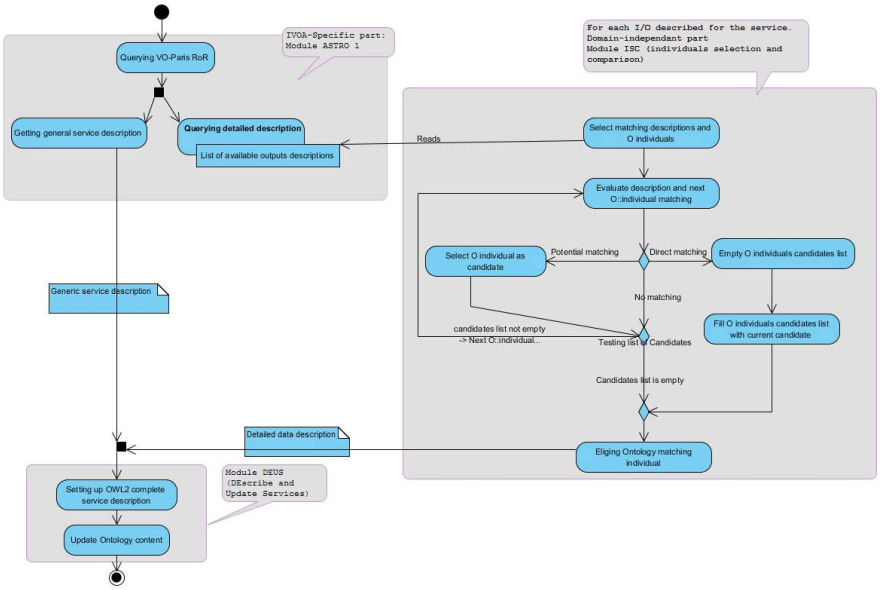

Fig. 4. Information matching detection

For the use case exposed in 3, the user may give the request through natural langage, e.g. “Multiwavelength analysis of Crab nebula”, and give the system some more informations by using a web interface to specify some more informations, as coordinates of the target, specific wavelenghts to ignore or to priorize, etc. Combination of natural langage description and web interface specifications will lead to a request representation allowing to query the system to get every possible workflow and choose the more appropriate.

We’ll illustrate our system with this use case, saying that the user gives the system starting informations: target name (Crab nebula) and a radius (tolerance factor applied to object coordinates), and wants multiavelength analysis. The request representation matches those given informations with internal representation, being “multilambda” for result of multiwavelength analysis, “radius” and “target_name” for given object name and runs the system based on those requirements.

Generating a graph of possible workflows.

During the step of generating all paths, we will examine a basic workflows to determine if partial results are already available in previous compositions, and to determine their reusability, their enrichment and the necessary adaptations. We will use current methods of isomorphism search graph or subgraphs[4, 5], aiming at extracting workflows service states to reuse in a similar context, we will have to bring our own knowledge base in order to best use them. Among

these works, those studying the structure of workflows from Taverna [9]will be of great support.

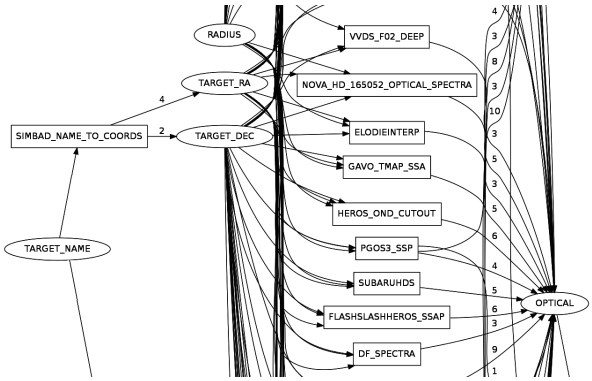

Fig. 5. Subsample of every possible path generated (testing purposes)

Figure 5, extracted from tests on the system, shows a subsample of the more than hundred possible paths generated by the system from informations given by the user leading to the multiwavelength analysis. Weights on the graph edges are randomized to elect best possible path during tests.

While the proposed system must be able to answer as many scientific problems posed, we must always be able to intervene in existing workflows to include the results of research from our own ontology. We must also, if abscence of oncoming already listed treatments, being able to explore the possibilities that we can offer independently.

4.4 Selecting one workflow.

Multiple workflows will be identified that lead to an answer of the user request, and a method is needed to identify the best workflow to choose. This requires information beyond the description of the problem, and may include past experience, preferred services, or preferred data sources. Any constraints or choices will be indicated explicitly to the user at all stages of the processing flow, and the user can modify these parameters to adapt the workflow selection.

Figure 6 illustrates best possible choice, based on random weights on services to obtain every needed information to go from informations given by the user to the result.

The given informations are used by the system that elect services (in squares) to provide informations (in rounds) to come to the final information, multiwavelength analysis.

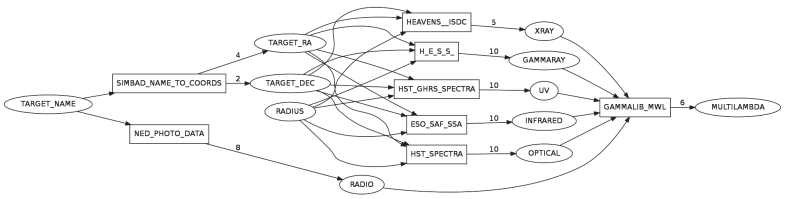

Fig. 6. One given path (generated for testing purposes)

5 Conclusion and future work.

We propose to describe astrophysical data and services using an ontology that connects these resources for arbitrary scientific workflows. Our system will rely on the Virtual Observatory initiative to ensure the interoperability of services although we also envision inclusion of non-VO services in our system. This work heavily relies on the use of ontological description of astrophysical quantities and services to cross-match generic, user-based descriptions of data and services with a structured knowledge of the domain. A few VO services use ontological description which matches with the notion of "Astroinformatics"[12]. This notion is related to the expanding number of data available and the need to provide useful and efficient tools to extract knowledge and sleeping science from this big data source. In our knowledge, nothing has been tried in this field using an ontological representation of knowledge as a base for automated service workflow discovery and composition from the description of a scientific use-case.

The challenge is to provide a good enough information recognition between services and requests from many different sources. It will allow the discovery of relevant services, and then organize them in order to produce results. Also, it allows to compare those results with other sources; as well as giving the possibility to the user to provide feedback and modify the entire workflow to fit very specific needs.

We still have to take into account some internal specificities of the services to be able to get a fully usable workflow and obtain complete results (corresponding to "execution layer" in Figure 2). Hence, we have to look at the need for services to obtain subset of informations they need through one service alone. There are cases when some subset of input informations (or all informations) for one service need to result of a unique source, others where such subsets may come from different sources and our system must be able to handle every case. Also, it will be necessary to work on the user-guided interface to express queries semantically uderstandable by the system. Actually, the use of randomize weights hasn't to be considered as the final goal. In our future works, we'll try to apply a more sophisticated method to choose accurate paths for every step of the flows.

References

1. M. G. Allen, F. Ochsenbein, S. Derrire, T. Boch, P. Fernique, and G . Landais; ; Astronomical Data Analysis Software and Systems XXIII ASP Conference Series, Vol. 485 N. Manset and P. Forshay, eds. c 2014 Astronomical Society of the Pacific; *Extracting Photometry Measurements from VizieR Catalogs.*
2. Robert Bentley, John Brooke, André Csillaghy, Donal Fellows, Anja Le Blanc, e Mauro Messerotti, David Pérez-Suarez,Gabriele Pierantoni, Marco Soldati; 2011 Seventh IEEE International Conference on eScience; *HELIO: Discovery and Analysis of Data in Heliophysics*
3. S. Derrière , R. W. Goosmann , C. Bot and F. Bonnarel; SF2A 2014; *Using the virtual observatory: multi-instrument, multi-wavelength study of high-energy sources*
4. Paolo Missier, Norman W. Paton et Khalid Belhajjame; *Fine-grained and efficient lineage querying of collection-based workflow provenance*; 13th International Conference on Extend- ing Database Technology (EDBT), 2010
5. D. Grigori, JC Corrales, M Bouzeghoub and A Gater; *Ranking BPEL Processes for Service Discovery;* IEEE transactions on service computing vol.3 Nb.3, July-September 2010
6. Mark D Wilkinson, Benjamin Vandervalk and Luke McCarthy; *The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern, API and Reference Implementation*; Journal of Biomedical Semantics 2011, 2:8
7. Taheriyan, M.; Knoblock, C. A.; Szekely, P.; and Ambite J .L.; *A Scalable Approach to Learn Semantic Models of Structured Sources*; Proceedings of the 8th IEEE International Conference on Semantic Computing, 2014.
8. Rahul Parundekar, Craig Knoblock, and Jose Luis Ambite*; Discovering Concept Coverings in Ontologies of Linked Data Sources*; Proceedings of the 11th International Semantic Web Conference (ISWC 2012), Boston, Massachusetts, 2012. Best Research Paper Award.
9. Jiuqiang CHEN; *Designing scientic workows following a structure and provenance -aware strategy* ; thèse de doctorat université Paris Sud; 2013
10. Laurino, Omar; Budynkiewicz, Jamie; D'Abrusco, Raffaele; Bonaventura, N.; Busko, I.; Cresitello-Dittmar, M.; Doe, S.; Ebert, R.; Evans, J.; Norris, P.; Pevunova, O.; Refsdal, B.; Thomas, B.; Thompson, R.; *Iris: An Extensible Application for Building and Analyzing Spectral Energy Distributions*; 2014, Astronomy & Computing, DOI: 10.1016/j.ascom.2014.07.004.
11. SONG Fuqi; Contribution of ontology alignement to enterprise interoperability; thèse de doctorat université Bordeaux1, 2013
12. Giuseppe Longo, Massimo Brescia, S.G. Djorgovski, Stefano Cavuoti, Ciro Donalek; *Data-driven dicovery in astrophysics*; astro-ph.IM; 21Oct2014
13. Alice Allen and Judy Schmidt. *Looking before leaping: Creating a software registry*. Technical Report 1407.5378, arXiv, 2014. http://arxiv.org/abs/1407.5378.
14. Tim Berners-Lee, James Hendler and Ora Lassila , *The Semantic Web: A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities*, Scientific American, May 2001.
15. J. Rao and X. Su, *A Survey of Automated Web Service Composition Methods*, Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition , 2005, pp. 43 – 54
16. B. Benatallah, M. Hacid, C. Rey and F. Toumani, *Semantic Reasoning for Web Services Discovery*, Proc. World Wide Web (WWW) Workshop E-Services and the Semantic Web, 2003.

17. Furkh Zeshan and Radziah Mohamad, *Semantic Web Service Composition Approaches: Overview and Limitations*, International Journal on New Computer Architectures and Their Applications, 2011.
18. Peter Bartalos , Mria Bielikov, *Automatic dynamic web service composition: A survey and problem formalization,* Computing and Informatics, Vol. 30, 2011