Web services as building blocks for Science Gateways in Astrophysics

Susana Sánchez Expósito*, Pablo Martín†, Jose Enrique Ruíz*, Lourdes Verdes-Montenegro*,
Julian Garrido*, Raül Sirvent‡, Antonio Ruíz Falcó† and Rosa Badia‡
*Instituto de Astrofísica de Andalucía - CSIC. Email: [sse,jer,lourdes,jgarrido]@iaa.es
†Fundación Centro de Supercomputación de Castilla y León. Email: [pmartin,antonio.ruizfalco]@fcsc.es
‡Barcelona Supercomputing Center. Email: [raul.sirvent,rosa.m.badia]@bsc.es

***Abstract*—An efficient exploitation of the Distributed Computing Infrastructures (DCIs) is specially needed to deal with the data deluge that the scientific community, in particular the Astrophysics one, is facing. This requires a good understanding of the underlying DCIs. Science Gateways (SGs) provide the users with an environment that eases the interaction with the DCIs. As a previous step, IT skilled users should populate the SGs with friendly but advanced tools (e.g. workflows, visualization tools) that not only support the scientists to build their own experiments but also adapt them in an optimal way to the infrastructures.**

In Astronomy, the Virtual Observatory provides the community with services and tools for data access and sharing. However, state of the art telescopes and the coming Square Kilometre Array (SKA), able to reach data rates in the exa-scale domain, will also require advanced tools for data analysis and visualization that should be run on DCIs as well as shared on SGs.

In the here presented work, we have selected as exemplar a set of analysis tasks of interest for some SKA use cases. These analysis tasks have been implemented as web services that use the COMPSs programming model in order to achieve a more efficient use of the DCIs. At the same time, the nature of the web services turns them into blocks that the astronomers can combine with VO services to build their own workflows. The web services and the workflows built upon them form a two-level workflow system that hides the technical details of the DCIs and exploits them efficiently.

This approach is used for the first time in analytical tasks of interest for the SKA that benefits from the capabilities of the DCIs.

***Index Terms*—Astronomy, Virtual Observatory, Web services, Workflows, Science Gateways, Distributed Computing Infrastructures**

I. Introduction

The development and use of distributed computing infrastructures (DCIs) has been mostly driven by the increasingly demanding scientific applications, which, paradoxically need to be adapted in order to be executed on these infrastructures. For an efficient use of them, scientists need to be aware of their technical characteristics [1]. This situation is far from ideal given the data deluge that the science community, and in particular the Astrophysics one, is facing.

The Science Gateways (SGs) provide the users with friendly tools that ease the interaction with the infrastructures. They allow the users to interact with the applications ported in the DCIs - browsing them, executing them, tracking their status and managing their input and output data -, create workflows built upon these applications, reuse workflows shared by other users, access advanced data browsing, and visualizing tools or portlets that ease the customization of the applications. In these environments, two user categories are distinguished (see e.g [2]): end-users and application developers. The latter are in charge of adapting the scientific software to the DCI and providing the former with the blocks to build their use cases. The application developers deal with scientific applications that in some cases are sequential and that miss hence the capabilities of the DCIs. Tools like COMPSs [3] or Dispel4py [4] facilitate the exploitation of the inherent parallelism of these applications. They are able to discover the dependencies among the application tasks, submitting them to the DCIs as much concurrently as possible and with a minimal transformation of the original code.

In MoSGrid [5], a Science Gateway based on WS-PGRADE [6], IT skilled users implement services and workflows for molecular simulations which can be reused by scientists to build their own experiments. VisIVO Science Gateway [7] gathers a set of workflows that run on DCIs for visualizing large-scale astrophysical datasets. It offers portlets to ease the parameter setting of the workflows, allowing the users to adapt them to their needs or use them as templates to build new ones. BioVEL project[1] exploits the advantages of the Software as a Service (SaaS) model, implementing a set of web services that interface several biological applications, that the end-users can combine in the Taverna Workbench [8]. The BioVEL users can share as well their workflows through the myExperiment portal[2] [9] or the BiodiversityCatalogue[3].

In Astronomy, the International Virtual Observatory Alliance IVOA[4] promotes the development of the Virtual Observatory (VO), a web-distributed interoperable data network using common standards for data publishing, discovery and sharing. The VO implements a distributed directory of information, called the VO Registry, where VO applications can discover and select resources (e.g. data archives and services) that are relevant for a particular scientific problem.

[1]https://www.biovel.eu/
[2]http://biovel.myexperiment.org/
[3]https://www.biodiversitycatalogue.org/
[4]http://www.ivoa.net

978-1-4673-7459-0/15/$31.00

This makes the development of VO Science Gateways easier, taking advantage of available standardised data collections and tools.

The astronomical community is preparing as well for the management of the heavy and complex datasets that will be generated by the SKA, an instrument that once built, will be the world's largest radio interferometer by far. It will be completed in its full extent around 2028 when it will be able to generate 10.000 petabytes per day. Therefore, SGs in Astronomy require high capacity computing and networking resources in order to empower VO-compliant data management services, as well as advanced tools for reduction and analysis of large datasets. The SKA pathfinder projects are working on building more efficient pipelines to reduce the data - i.e. to transform the stream of raw data into science ready data, applying Big Data techniques among others. An illustrative example is one of the projects [10] running LOFAR (LOw Frequency ARray) radiotelescope. It stores the data in a HPC cluster, specifically designed towards data intensive work, and accessed through a MonetDB database that accelerates the automated pipeline aiming to detect transient sources. However, an effort is still needed to improve the tools for analysing and visualizing large volumes of science ready data. They should as well be incorporated into environments where a wide user community can access them. Along this line, we find CyberSKA [11], a collaborative web portal for radio astronomy that provides access to data management and visualization tools.

In the here presented work we have focused on those SG blocks dedicated to the analysis of the data, being data management as well as pipelines for data reduction beyond the scope of this paper. In particular, we have developed a collection of services for the analysis of interferometric data. In order to drive the development of these services, we have selected as exemplar a set of analysis tasks of interest for those SKA use cases aiming to study given parameters of galaxies (see Sect. II for details). We refer to them as AMIGA4GAS services since they have been developed within the framework of the AMIGA for GTC, ALMA, and SKA pathfinders project[5] (AYA2011-30491-C02), whose technical objective is to build a federated layer that facilitates astronomers launching their workflows in heterogeneous DCIs. The AMIGA4GAS services have been internally implemented with COMPSs programming model in order to achieve a more efficient use of the DCIs, while externally their web interface turns them into blocks that the astronomers can combine with VO services to build their own workflows. In this paper we present this two-level workflow system: at the user level the workflows are built upon web services, while those services turn to be as well workflows at the infrastructure level, and are built as COMPSs applications. Next section depicts the science use case. Then, Sect. III describes the two-level workflow system and Sect. IV provides a list of the implemented services and workflows. The main conclusions of the exposed work are presented in Sect. V.

[5]http://amiga.iaa.es/p/263-federated-computing.htm

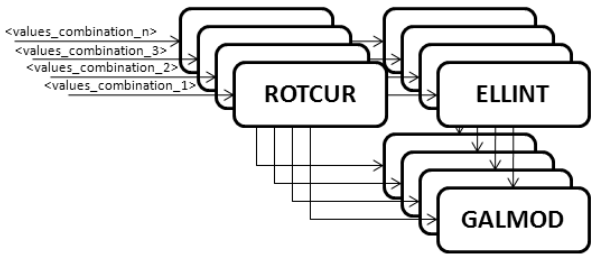

Fig. 1. Use case schema

II. Science Use Case

One of the SKA science drivers is the study of the galaxy evolution, for which the analysis of the atomic gas (HI) is required. Radio interferometers as the SKA generate data with two spatial coordinate axes and a spectral axis containing information about the atomic gas distribution and velocity. The astronomers analyse these HI datacubes to produce a kinematical model of the galaxies, key to understand the mass distribution, track possible perturbations due to different phenomena or define the dynamical structure and hence the distribution of matter in galaxies.

The Groningen Image Processing System (GIPSY) [12] is a powerful software package specially suited for this kind of kinematical analysis, via mainly three GIPSY tasks: ROTCUR, ELLINT and GALMOD. ROTCUR derives the kinematical parameters from the datacube, including the rotation curve; ELLINT generates the radial profile of the galaxy emission; GALMOD builds a model from the kinematical parameters and the emission profile.

The users provide a set of input data some of them compiled from external catalogues, or derived by visual inspection of the data. It is a common practice to test different values of them, sweeping a small range, to generate several models. Then the astronomer assess which model better fits the original data. This use case is represented in Figure 1 where is shown both the connection among the tasks and the needed parallelism due to the range of input values.

III. Two-level workflow system design

The two-level workflow system consists of a set of web services implemented as COMPSs applications and a set of Taverna workflows built upon these web services. The former are considered as workflows at low level, close to the infrastructures, since COMPSs orchestrates the internal tasks of the services in order to submit them to the DCIs as much concurrently as possible. We refer to these services as AMIGA4GAS services. They have been implemented by IT skilled users in charge of porting the GIPSY package to the computing infrastructures, implementing the COMPSs applications and designing a web interface that allows the astronomers to access the ported GIPSY tasks from a friendly environment like Taverna Workbench.

The high level workflows, close to the users, are the Taverna workflows built upon the AMIGA4GAS services. The

astronomers can build them using the AMIGA4GAS services as blocks that can be combined in different ways to customize an experiment.

Next subsections describe in detail both levels of the system design.

A. Infrastructure Level: Workflows as Web Services

1) Architecture: The AMIGA4GAS services are described with the Web Services Description Language (WSDL) and are transmitted using the Simple Object Access Protocol (SOAP). They have been designed with different granularity levels (see Sect. III-A3) ranging from services interfacing a single GIPSY task to services implementing the whole analysis process. Internally, these services call COMPSs applications that orchestrate the tasks involved in the services and submit them to the DCIs. Fig. 2 shows the WSDL/SOAP interface on top of the COMPSs framework that manage the interaction with the computing infrastructures where GIPSY is installed.

COMPSs is a programming framework that exploits the inherent parallelism of the applications. In this framework the users mark the methods invoked from their sequential application to be run as remote tasks on the available resources. COMPSs, at execution time, discovers the dependencies among these tasks, launching them to the computing resources as soon as their input data are ready. So, COMPSs is able to parallelize those pieces of code of the sequential application that can be executed concurrently. COMPSs uses JavaGAT connectors to launch jobs to different infrastructures, like grid infrastructures based on gLite and Sun Grid Engine clusters. It has as well an Open Cloud Computing Interface (OCCI) connector that enables COMPSs to submit jobs to those cloud infrastructures supporting this interface. Therefore the COMPSs applications not only orchestrate the service tasks while exploiting their parallelism, but also act as infrastructure broker, submitting the tasks to the DCIs, checking their execution status and gathering their outputs.

We have as well ported GIPSY to IBERGRID, the joint Spanish and Portuguese Grid infrastructure, and it is available for the users of the Virtual Organization phys.vo.ibergrid.eu. It has been installed as well in the supercomputing cluster at the Fundación de Supercomputación de Castilla y León (FCSCL), enabling thus that the COMPSs applications submit the tasks services to both infrastructures.

2) User Interface design: The AMIGA4GAS services act as interface of one or several GIPSY tasks. Those tasks have a large number of input parameters allowing them to perform several kind of operations. Handling this amount of parameters in a graphical workflow manager as Taverna can be tedious. Reducing the number of input parameters by setting some of them with a default value constrains the specific operation that the service will perform. This would help non experienced GIPSY users, since they do not get confused by those parameters that they do not need. However, expert GIPSY users could miss some functionalities. So, once the input parameters related with a graphical/interactive mode of the tasks were discarded, we took the decision to focus the services to the operations related to the kinematical modelling of galaxies, trying to find a compromise between flexibility and usability. On demand, we could incorporate new services that would implement more advanced operations to fulfil the requirements coming from expert users.

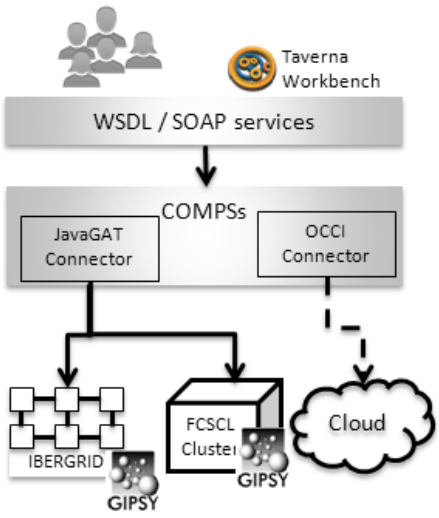

Fig. 2. Architecture of the AMIGA4GAS services

Astronomers usually explore different combinations of values for certain input parameters that are difficult to estimate with accuracy. This repetitive and time consuming job should be automated to cope with the large amount of expected data. Therefore, some AMIGA4GAS services have been implemented to receive a range or a list of values for these parameters and execute the specific GIPSY task as many times as the possible combinations. Notice that the user can also build a workflow in Taverna that iterates over a list of values, calling the service the needed times to sweep the values range. However the time execution of this workflow will be higher since the latency time of invoking the web service plus the time of receiving its output will slow down the whole execution.

3) Granularity: The web services have been implemented with different granularity levels aiming to be adaptable enough to the different user needs, ranging from services interfacing a single GIPSY task to services interfacing the three GIPSY tasks involved in the use case. The former configuration allows the user to combine the services in different ways, providing more flexibility to build new experiments. The execution time of the latter will be lower because 1) the communication among the tasks happens inside the web server, instead of between the user desktop (where Taverna is running) and the web server; and 2) the latter benefits from COMPSs' orchestration capability.

B. User Level: Taverna Workflows

The AMIGA4GAS services have been specially customized for the Taverna Workbench Astronomy 2.5, a special edition of the Taverna Workbench that includes support for building and executing astronomy workflows based on Virtual Observatory

TABLE I
LIST OF THE MAIN AMIGA4GAS SERVICES

WS Name	GIPSY Tasks	Outputs
rotcur_ws	ROTCUR	Parameters describing the rotation curve of the galaxy
ellint_ws	ELLINT	Radial profile of the galaxy emission
galmod_ws	GALMOD	A datacube modeling the input one
rotellint_ws	ROTCUR ELLINT	Kinematical parameters and emission radial profile
modeling_ws	ROTCUR ELLINT GALMOD	Kinematical parameters, emission radial profile and a datacube model

TABLE II
LIST OF TAVERNA WORKFLOWS

Wf Name	WS involved	myExp. ID
rotcur_wf	rotcur_ws	4609
rotellint_wf	rotellint_es	4611
rotcur_ellint_wf	rotcur_ws, ellint_ws	4610
modelling_wf	modelling_ws	4613
rotcur_ellint_galmod_wf	rotcur_ws, ellint_ws, galmod_ws	4612
VO_rotcur_wf	VO services and rotcur_ws	4619
VO_modelling_wf	VO services and modelling_ws	4620

(VO) through the Astrotaverna plugin [13]. This plugin integrates existing VO web services as first-class building blocks in Taverna workflows as well as it provides this workflow manager with astronomical data manipulation tools.

IV. IMPLEMENTED SERVICES AND WORKFLOWS FOR THE KINEMATICAL MODELLING OF GALAXIES

The AMIGA4GAS services can be imported in Taverna Workbench or other WSDL clients, through a specific URL[6]. The access to the services is authenticated, hence an user login and password should be provided. The AMIGA4GAS services website[7] contains detailed description of the services as well as an user manual.

Table I lists the main AMIGA4GAS services. The services rotcur_ws, rotellint_ws and modeling_ws admit a range or list of input values, obtaining as many outputs as possible combinations of these values.

A set of Taverna workflows has been as well implemented using the AMIGA4GAS services. Table II lists the web services involved in each workflow together with the myExperiment ID that can be used to access the workflow[8].

The workflows rotellint_wf and rotcur_ellint_wf have the same functionality. The difference between them is the use of services with different granularity. While the former uses a single service interfacing two GIPSY tasks, the latter uses two services, each one representing one single GIPSY task. The same applies to modelling_wf and rotcur_ellint_galmod_wf.

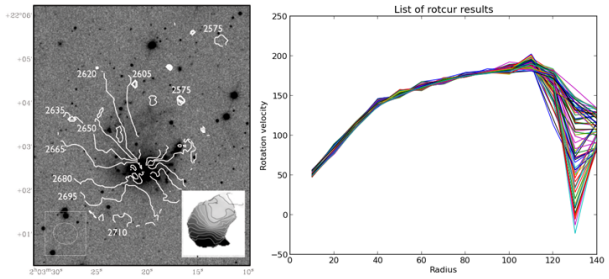

Fig. 3. HI velocity field contours of the CIG85 galaxy and the results of exploring an input parameter range to calculate the rotation curve of this galaxy (rotcur_wf)

The VO_rotcur_wf and VO_modelling_wf workflows integrate a subworkflow that returns the location of the datacube that being analysed. This subworkflow is composed of two Virtual Observatory compliant web services: SIAv2 and DataLink. While the former implements an upgraded version of the IVOA Simple Image Access[9] protocol for the discovery of HI datacubes, the latter implements the IVOA DataLink[10] protocol to provide additional information. Both protocols are in the state of Proposed Recommendation in the IVOA Data Access Layer Working Group and in their way to reach the Final Recommendation status as *de facto* standards. The SIAv2 service queries a particular datacube archive, searching for datacubes inside a sky region specified by the user through e.g. a pair of coordinates and a search radius. Once the datacube/s fulfilling the search criteria are identified, the DataLink service delivers related complementary metadata, including the datacube/s location. Until now, we have adopted two formats to define a location. In the case of datacubes stored in clusters, their location is represented as the server name followed by *@:* symbols and the local path of the dataset (e.g. srv-prj-wsamiga.fcsc.es@:/Amiga4GAS/3DCIG0232/CIG0232), whereas the Logical File Name (LFN) is used in case of datacubes in gLite grids (e.g. lfn:/grid/phys.vo.ibergrid.eu/jgarrido/whisp/3DCIG0232).

In Fig. 3 we present an example illustrating the selected science case. The left panel shows an optical image of a galaxy (CIG85 [14]) with HI velocity field contours. Those contours are used to get some of the input values required by rotcur_wf workflow. The result of executing it is a set of 81 rotation curves showed at the right hand panel of the figure.

V. CONCLUSIONS

The scientific communities are approaching the data deluge challenge through Science Gateways where the scientists can

[6]https://srv-prj-wsamiga.fcsc.es:8444/Amiga4GasServiceLadon/soap11/description
[7]https://srv-prj-wsamiga.fcsc.es/
[8]http://www.myexperiment.org/workflows/{ID}.html
[9]http://www.ivoa.net/documents/SIA/
[10]http://www.ivoa.net/documents/DataLink/

share and reuse tools that exploit efficiently the DCIs for analysing and visualizing large data sets. The astronomical community is paving the way to the SKA, an instrument that once built, will be the world's largest radio interferometer by far, able to reach data rates in the exa-scale. Big data techniques are being applied to improve the pipelines for the SKA data processing. However an effort is still needed in this field to improve the science ready data analysis, through advanced services that both exploit efficiently the DCIs and keep the astronomers unaware of the technical complexity of them.

This work aims to contribute to these efforts with a two-level workflow system that at low level uses tools like COMPSs to make a good use of the computing resources, and at high level provides a web interface to facilitate the astronomers to build their workflows. This system has been applied to build a set of advanced tools that implement analysis tasks of interest for those SKA use cases aiming to study given parameters of galaxies. At infrastructure level, we designed a set of web services with different level of granularity to be adapted to different user preferences. These benefit from COMPSs application model to exploit efficiently the computing infrastructures. At user level, we developed a set of Taverna workflows built upon the previous web services. They implement different use cases that can be used as templates, including workflows that combine kinematical analysis services with Virtual Observatory services.

This work constitutes the first application of the two-levels model used in other disciplines to Astronomy, and in particular to tasks of interest for SKA that benefit from the capabilities of the DCIs. It has been implemented in a way that the technical details at the lowest level are transparent for the astronomer.

Acknowledgement

This work has been developed in the framework of AMIGA for GTC, ALMA, and SKA pathfinders project, funded under AYA2011-30491-C02, co-funded by MICINN and FEDER funds. This work has been also supported by the Wf4Ever Project[11] 270129 funded under EU FP7Digital Libraries and Digital Preservation (ICT-2009.4.1) and the Galaxias y Cosmologia project funded by the Junta de Andalucia (TIC-114).

This work as well has benefited from the IBERGRID infrastructure.

References

[1] T. Kiss, I. Kelley, and P. Kacsuk, "Porting computation and data intensive applications to distributed computing infrastructures incorporating desktop grids," in *Proceedings of Science*, 2011.

[2] "Science gateway primer," EGI, Tech. Rep., 2013.

[3] F. Lordan, E. Tejedor, J. Ejarque, R. Rafanell, F. M. Javier lvarez, D. Lezzi, R. Sirvent, D. Talia, and R. M. Badia, "Servicess: An interoperable programming framework for the cloud," *Journal of Grid Computing*, vol. 12, no. 1, pp. 67–91, 2014.

[4] R. Filguiera, I. Klampanos, A. Krause, M. David, A. Moreno, and M. Atkinson, "Dispel4py: A python framework for data-intensive scientific computing," in *Proceedings of the 2014 International Workshop on Data Intensive Scalable Computing Systems*, ser. DISCS '14. Piscataway, NJ, USA: IEEE Press, 2014, pp. 9–16.

[5] J. Krüger, R. Grunzke, S. Gesing, S. Breuers, A. Brinkmann, L. de la Garza, O. Kohlbacher, M. Kruse, W. E. Nagel, L. Packschies, R. Müller-Pfefferkorn, P. Schäfer, C. Schärfe, T. Steinke, T. Schlemmer, K. D. Warzecha, A. Zink, and S. Herres-Pawlis, "The mosgrid science gateway – a complete solution for molecular simulations," *J. Chem. Theory Comput.*, vol. 10, no. 6, pp. 2232–2245, 2014.

[6] A. Balasko, F. Z, and K. P., "Building science gateway by utilizing the generic ws-pgrade/guse workflow system," *Computer Science*, vol. 14, no. 2, p. 307, 2013.

[7] E. Sciacca, M. Bandieramonte, U. Becciani, A. Costa, M. Krokos, P. Massimino, C. Petta, C. Pistagna, S. Riggi, and F. Vitello, "] visivo science gateway: a collaborative environment for the astrophysics community," in *International Workshop on Science Gateways*, 2013.

[8] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, J. Bhagat, K. Belhajjame, F. Bacall, A. Hardisty, A. Nieva de la Hidalga, M. P. Balcazar Vargas, S. Sufi, and C. Goble, "The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud," *Nucleic Acids Research*, vol. 41, no. W1, pp. W557–W561, 2013.

[9] D. De Roure, C. Goble, and R. Stevens, "The design and realisation of the myexperiment virtual research environment for social sharing of workflows." *Future Generation Computer Systems*, vol. 25, pp. 561–567, 2009.

[10] B. Scheers and F. Groffen, "Towards dynamic light-curve catalogues," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 8451, Sep. 2012, p. 0.

[11] C. Kiddle, A. R. Taylor, D. Pigat, O. Eymere, E. Rosolowsky, V. Kaspi, and A. G. Willis, "CyberSKA : An On-line Collaborative Portal for Data-intensive Radio Astronomy," in *ACM workshop on Gateway computing environments*, 2011, pp. 65–72.

[12] M. G. R. Vogelaar and J. P. Terlouw, "The evolution of gipsy, or the survival of an image processing system," in *Astronomical Data Analysis Software and Systems X*, vol. 238, 2001, p. 358.

[13] J. Ruiz, J. Garrido, J. Santander-Vela, S. Sánchez-Expósito, and L. Verdes-Montenegro, "Astrotaverna - building workflows with virtual observatory services," *Astronomy and Computing*, vol. I, pp. 3–11, 2014.

[14] C. Sengupta, T. C. Scott, L. Verdes Montenegro, A. Bosma, S. Verley, J. M. Vilchez, A. Durbala, M. Fernández Lorenzo, D. Espada, M. S. Yun, E. Athanassoula, J. Sulentic, and A. Portas, "H I asymmetry in the isolated galaxy CIG 85 (UGC 1547)," *Astronomy and Astrophysics*, vol. 546, p. A95, Oct. 2012.

[11] http://www.wf4ever-project.org/