2014 IEEE 10th International Conference on eScience

Scientific Workflow Management - for Whom?

S.D. Olabarriaga, M. Jaghoori
Academic Medical Center
Univ of Amsterdam, NL

G. Taffoni, G. Castelli, C. Vuerli
INAF Astronomical Observatory
Trieste, IT

G. Pierantoni, E. Carley
Trinity College Dublin
Dublin, IR

V. Korkhov
St.Petersburg State University
Russia

E. Sciacca, U. Becciani
INAF Astrophysical Observatory
Catania, IT

B. Bentley
Mullard Space Science Laboratory
London, UK

***Abstract*—Workflow management has been widely adopted by scientific communities as a valuable tool to carry out complex experiments. It allows for the possibility to perform computations for data analysis and simulations, whereas hiding details of the complex infrastructures underneath. There are many workflow management systems that offer a large variety of generic services to coordinate the execution of workflows. Nowadays, there is a trend to extend the functionality of workflow management systems to cover all possible requirements that may arise from a user community. However, there are multiple scenarios for usage of workflow systems, involving various actors that require different services to be supported by these systems. In this paper we reflect about the usage scenarios of scientific workflow management based on the practical experience of heavy users of workflow technology from communities in three scientific domains: Astrophysics, Heliophysics and Biomedicine. We discuss the requirements regarding services and information to be provided by the workflow management system for each usage profile, and illustrate how these requirements are fulfilled by the tools these communities currently adopt. This paper contributes to the understanding of properties of future workflow management systems that are important to increase their adoption in a large variety of usage scenarios.**

I. Introduction

Scientific workflow management systems (SWMS's) have become valuable tools to carry out complex scientific experiments. They offer means to compose and distribute steps needed to perform computations for data analysis and simulations, whereas hiding details about the complex infrastructures underneath. More importantly, workflow descriptions capture the process of scientific experimentation, which are useful to reproduce, reuse or re-purpose these processes.

Active research in scientific workflow management has resulted in a large number of systems that can be used by scientists in practice. These systems address a large variety of scientist's needs. They offer generic services to handle distribution, monitoring and fault-tolerant distributed computations in various types of platforms (e.g. web services, grids and clouds); automatic capture of provenance of the involved processes and data; workflow composition and progress monitoring; etc. When thinking about "users" of SWMS's and their requirements, one typically has in mind an ideal scientist, who is interested in the complete functional chain of workflow management. In the daily practice of various user communities this is simply not the case. Workflow systems, as powerful tools as they are, can be deployed in multiple scenarios and serve the needs of various persons with complementary roles in the scientific experimentation chain. This gives rise to multiple profiles of "users" and, accordingly, diverse requirements for workflow management.

In this paper we reflect about usage scenarios and requirements for SWMS's based on the practical experience of heavy users of workflow technology from communities in three scientific domains: Astrophysics, Heliophysics and Biomedicine. Although the scientific applications and culture of these three communities are very diverse, they all adopt science gateways powered by a SWMS to distribute computation on various types of computing infrastructures. The ER-flow[1] and the SCI-BUS[2] projects provide the context for the collaboration that motivated this study. The ER-flow project aims at building a workflow user community across Europe, providing an opportunity to exchange experiences across scientific domains around the topic of scientific workflows. The project exploits the SHIWA[3] workflow interoperability platform, which grants workflow developers the freedom to choose their preferred workflow system for development, whereas enabling the execution of all these workflows expressed in different languages within the same system. The SCI-BUS project aims at creating a new science gateway customization methodology based on the generic-purpose gUSE/WS-PGRADE [13] portal family. A science portal or gateway is defined here as a community-developed set of tools, applications, and data that are integrated via a portal or a suite of applications with graphical interfaces that are further customized to meet the needs of a targeted community. The computational processes supported by the science gateway are organized as scientific workflows that specify dependencies among underlying tasks for orchestrating distributed resources (such as clusters, grids or clouds). Science gateway technologies such as gUSE/WS-PGRADE allow the scientific research community to create a web-based working environment where researchers can concentrate on scientific problems without facing the complexities of the computing, data and workflow infrastructure.

During our collaborations we identified similarities in the patterns of usage of workflow systems across these communities. These patterns reveal that, besides the known requirements for supporting scientist users of a SWMS, additional ones emerge to support the development and operation of science

[1] http://www.erflow.eu
[2] http://www.sci-bus.eu
[3] http://www.shiwa-workflow.eu

978-1-4799-4287-9/14 $31.00 © 2014 IEEE
DOI 10.1109/eScience.2014.8

gateways that are based on the workflow system. Understanding patterns and requirements of these particular cases is important to indicate future directions of further improvements of SWMS's.

The paper is organized as follows. Related works in requirements for SWMS's are presented in Section II. The background of each user community is presented in Section III. Section IV analyses the various usage scenarios based on the user goals and profiles and workflows. Next it discusses the requirements regarding services and information to be provided by the SWMS for each usage profile, illustrating how these requirements are fulfilled by the tools these communities currently adopt. Section V and VI close the paper with discussions and conclusions. This structured analysis of users, profiles and requirements contributes to understanding properties of future SWMS's that are important to increase the adoption of this technology in a variety of usage scenarios.

II. Related work

A large amount of literature is devoted to understanding properties and requirements of SWMS. An early work [30] proposes a generic architecture for grid workflow systems that is based on the workflow reference model defined by the Workflow Management Coalition [28]. The model separates between workflow definition (build time) and workflow execution (run time). It also defines the main components and functionalities of a workflow management system: workflow definition, workflow specification, grid resources, information services, and a workflow enactment engine capable of scheduling, data transfers and fault tolerance. This model however leaves out the layer "interaction with user and application tools", which was defined by the original reference model. From the proposed architecture, users interact with the workflow system for workflow design (build time), giving the impression that workflow execution (run time) is fully automatic.

The results of a workshop organized in 2006 present the challenges for large scale adoption of workflows in science [29]. Various levels are highlighted: application, workflow description and evolution, and system-level workflow management. This work describes many important requirements, in particular the need for flexible environments with interfaces for the various services, to address different user needs. Here "users" are the scientists who want to perform scientific experiments using workflows.

Suggestions and recommendations for future development of workflow management systems are also presented in [2], in particular for the interfacing with users. We quote: "the workflow system should allow the same information to be shown at various levels of abstraction depending on who is using the system". This work also emphasizes the need for customized portal-based access and scripting interfaces as means to address usability and flexibility needs.

Deelman et al. [9] have further characterized the features of workflow systems. They divide the life-cycle of scientific workflow management into four phases: workflow composition, mapping onto resources, workflow execution and provenance capture. The capabilities to support these four phases are discussed with examples from existing systems. This work has introduced considerations about workflow provenance and interoperability. It also takes into account the perspective of a scientist and distributed application developer, however without clearly identifying different needs for each one.

The more recent work of Cerezo et al. [7] revisits the concepts in workflow management from an accessibility perspective, disentangling the large and complex aspects involved in a SWMS. The work identifies three levels of workflow abstraction: Concrete Level (actual execution on a particular DCI), Abstract Level (ready to be interpreted or compiled, but not entirely bound to specific resources); and Conceptual Level (at which scientists conceive scientific experiments in a vocabulary that is familiar to them). These levels represent different information about scientific workflows, being also useful to understand the needs of users that might be interested in only a sub-set of them. Cerezo also lists in [8] the various types of user interfaces to workflow management systems: application programming interface (API), command-line, graphical, portal, file formats, scripting and webservices. However, also in the work of Cerezo the user of workflow systems is pictured as a domain scientist.

All these works recognize the large variety of scientist users with different profiles, and accordingly the diversity of requirements. They however tend to ignore other actors that are implicitly involved in the development and operation of sophisticated virtual research environments, and who also need to communicate with the SWMS via some kind of interface.

III. Studied Scientific Communities

In this section we briefly present how three scientific communities use SWMS to support their research. The following aspects are covered: background of the scientific area, e-infrastructure, interfaces (science gateways) between the scientists and infrastructure, and people involved.

A. Astrophysics

Astronomy is a natural science that deals with the study of celestial objects, and Astrophysics is the branch of astronomy that deals with the physics of the Universe, including physical properties of celestial objects, their interactions and behaviour. Astrophysics has become a data intensive science due to numerous digital sky surveys, with many TB of pixels and with billions of detected sources, and often with tens of measured parameters for each object. Moreover, high-resolution numerical simulation codes are producing in-silico experiments that result in PB of data to be stored and analyzed. Handling and exploring these new data volumes, and actually making real scientific discoveries, pose a considerable technical challenge that needs to overcome the traditional research methods in these sciences. e-Infrastructures provide a vital foundation for the Astrophysics community, such as the European Grid Infrastructure (EGI[4]) and the Open Science Grid[5]. In particular the Virtual Observatory data infrastructure of the International Virtual Observatory Alliance (IVOA[6]) offers tools, software and services to access, share, manipulate and visualise data.

Workflow systems have been widely used to coordinate services and to access computing resource and data storage.

[4]http://www.egi.eu
[5]www.opensciencegrid.org
[6]http://www.ivoa.net

For example, in the Workflow4Ever project[7], the Astrophysics community has developed more than 50 workflows using Taverna [15] and the AstroTaverna plugin [20]. AstroTaverna integrates existing Virtual Observatory web services as first-class building blocks in Taverna workflows (e.g. to search a registry, add found services to the workflow, manipulate data in form of VOTables, and convert coordinates). The AstroTaverna workflows resulted from a successful cooperation among astronomers, who provide requirements and use the workflows, and computer scientists, who design and develop the workflows. *Data-oriented* workflows are used to interact with data, being mainly designed to search and get data in distributed database systems, manipulate data or perform simple data analysis tasks. Each of these "Atomic" operations are implemented by an individual simple workflow. These workflows are simple to operate and do not demand large computing effort, so they are executed using the Taverna Desktop environment, by an astronomer. The data tasks run locally or on clusters using IVOA standards to access computing resources. Another class of workflows are *visualization-oriented*. They usually require computational demanding jobs to import, filter, extract useful metadata and visualise the datasets on DCIs to obtain meaningful information from the dataset. As parameters are varied within user-defined ranges, several hundreds to thousands of workflow executions might be necessary. For example, the creation of a movie represents a significant challenge for the underlying computational resources, as often hundreds or thousands of high quality images must be produced. Parameter sweep workflows are employed for visualization-oriented workflows, and they also can be used as building blocks in more complex workflows. Finally, there are also *computing-oriented* workflows consisting of computing tasks. In this case we identify two different workflow patterns. The first pattern involves running multiple instances of the same workflow on different inputs, exploring different parameters. The second pattern consists of analysing different data using the same workflow, which is the case of data reduction/analysis pipelines. Such workflows are generally complex to design and implement, and for this reason they are developed by computer scientists based on requirements identified by astronomers. The researchers use the workflows thanks to simplified interfaces of science gateways.

During the SCI-BUS and ER-flow projects the Astrophysics community has gained experience in workflow design and implementation [3]. Since most of the astronomers were not familiar with workflow technologies, workflow developers have provided a set of core-workflows that can be easily set-up and submitted through a dedicated science gateway. The astronomers actually interact with the science gateway, while computer scientists are in charge of installing and maintaining the gateway, designing and deploying the workflows, and designing the user interfaces. Thanks to data and visualisation building blocks, some astronomers and astrophysicists have developed a special interest in workflow technology. They learned how to reuse the building blocks to create their own applications or combine workflows into meta-workflows.

The science gateways available for this community are implemented with gUSE/WS-PGRADE. One example is VisIVO, which allows scientific visualisation and analysis of large-scale multidimensional datasets [22][21], also from mobile devices [26], employing visualization-oriented workflows. Another example is the Planck science gateway[8] [19], which allows researchers to run parameter sweep simulations on a gLite Grid [25] from a simple front-end. Figure 1 shows an overview of the infrastructure related to Astrophysics community and in particular to VisIVO science gateway. The astrophysicists access the science gateway and mobile application, which rely on WS-PGRADE to connect to several DCIs. Furthermore, thanks to the workflow interoperability solution provided by the SHIWA technology, access to the Virtual Observatory services directly from Taverna is also enabled.

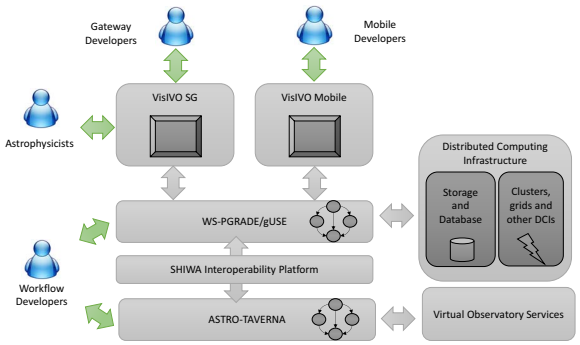

Fig. 1. Infrastructure for Astrophysics research

Recently a federation of Astrophysics-oriented science gateways, named STARnet, has been designed and implemented [4]. STARnet is based on gUSE/WS-PGRADE and it envisages sharing services for authentication, a common and distributed computing infrastructure (clusters or DCIs), data archives and workflow repositories. The first implementation of STARnet provides workflows for cosmological simulations, data post-processing and scientific visualization.

B. Heliophysics

Heliophysics investigates the interactions between the Sun and the other bodies of the Solar System. Heliophysicists use data collected by satellites, telescopes and other instruments to study events such as Coronal Mass Ejections (CMEs) that originate in the Sun and effect the other planets. Raw data is calibrated and processed to extract metadata that describe relevant features of the various events. This metadata is then used to build indexes of features and events called metadata catalogues. Heliophysicists perform three main kinds of processing activities: metadata extraction, metadata analysis and the modeling of solar events. Raw data calibration and metadata extraction are usually computationally and data intensive tasks, and require significant distributed resources. The analysis of metadata, on the other hand, usually requires to orchestrate queries on multiple and distributed sources. In addition to the data analysis and metadata query, heliophysicists develop conceptual and mathematical models of the phenomena and the environment of the Solar System, and test them against the scientific evidence gathered so far. A pressing issue is the need to simulate and understand how phenomena propagate throughout the Solar System. Scientists tackle this with mathematical

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

[8]Planck was an ESA Space mission aimed at mapping the microwave sky

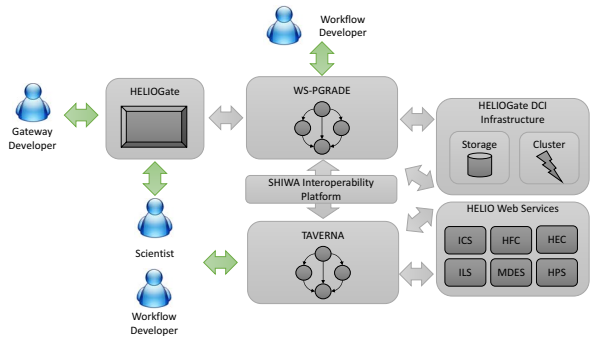

Fig. 2. Infrastructure for Heliophysics research

and physical abstractions called Propagation Models, for which numerous numeric tools exist. Data processing tools need large DCIs and the applications can be used for sustained periods, offering results to many different users. On the other hand, metadata query orchestration for the investigation of events is not usually data or computation intensive, but rather require high flexibility. The latter usage scenarios may be relevant for shorter periods of time and to a smaller number of scientists. In extreme but not uncommon cases, a metadata investigation may be relevant only to one scientist to investigate a specific event, and then discarded. Finally, event models can be computationally intensive and also require flexibility, as different scientists need to adapt the model to the needs of their current investigation. The HELIO project[9] [5] [17] built a Virtual Observatory for Heliophysics capable of integrating multiple data sources within an unified framework. HELIO also introduced workflows in Heliophysics [6], recognizing how promising this technology was for this field. The HELIO project left a legacy of Web Services and a first set of Taverna workflows for orchestration purposes. This legacy has been further developed during the SCI-BUS and ER-flow projects, which are building user interfaces for scientists and support for the execution of parameter-sweep applications over multiple DCIs.

Figure 2 presents an overview of the current infrastructure for Heliophysics. In addition to the traditional DCI infrastructure, a set of web services (legacy of the HELIO project) offer a standardized interface based on the IVOA standard for the query of different metadata catalogues. The community relies on two different SWMS's, WS-PGRADE and Taverna, which can interoperate through the SHIWA interoperability platform. The HELIOGate portal offers an interface for domain experts that want to be shielded from the complexities of the workflows.

Scientists here can be roughly divided into two broad categories: those solely interested in the results of the investigation, and those who share an interest in the workflow technology. The first scientists want to be shielded as much as possible from the implementation details of the workflows, being best served by dedicated graphical user interfaces that hide all the technical details. The second user type instead is interested in the underlying technologies and willing to be involved in the design, modification and execution of workflows. These scientists are likely to use the dedicated user interfaces for the execution of their workflows, but will also use the workflow editing and submission interfaces directly. There also are developers of workflows and portlets for the portal, who have a different background and purpose, although significant overlapping can occur. Their interest is more focused on the enabling technology (debugging and logging, API, etc.) rather than on the scientific results. The scenario is further enriched by the multiple workflow technologies adopted by the community. Most of the scientists use Taverna as their prime choice (Taverna is particularly suited for web services orchestration), while WS-PGRADE offers rapid prototyping of user interfaces and strong support for parameter-sweep jobs (relevant for data processing and statistical analysis of multiple events). The two technologies are bridged by the SHIWA interoperability platform, which allows for execution of Taverna workflows from a customized user interface of WS-PGRADE.

C. Biomedicine

Biomedicine is a branch of medical science that applies biological and other natural-science principles to clinical practice. This sub-field of life sciences has the aim of understanding the mechanisms of diseases, how they manifest themselves in detectable ways, and how they can be influenced to treat the patient. A large variety of resources are used in biomedical research, including data collections and analysis, simulation, modelling and experimentation, both in-vivo and in-vitro. The Academic Medical Center of the University of Amsterdam (AMC) is an active player in biomedical research, engaging a large community of biomedical scientists who carry out research mostly based on the analysis of large data collections of various types. Some examples of data that are daily used in our organization are medical images generated from various scanning modalities, genomics data from various sources, models and simulations. Typically, local workstations are used by AMC researchers to carry out their experiments. With the rapid growth of data variety, quantity and complexity, e-infrastructures have become important means to address modern biomedical research problems in our organization.

The Dutch e-science infrastructure comprises grid, cloud, storage and other resources that can be exploited for biomedical research. Nevertheless, their usage remains difficult for bioscientists with the usual low-level interfaces offered by these infrastructures. Therefore, there is a large effort on-going at the AMC to build customized, high level user interfaces that enable the scientists themselves to use this infrastructure. These science gateways are developed and operated by the e-science group of the AMC, which has computer science and engineering background. Currently the gateways are used by researchers from three main areas: neuroscience, biochemistry and genomics, in particular for next generation sequencing. The science gateway takes care of details such as data movement between the data and computation resources, collection of provenance information about the experiments, and executing/monitoring computations. The computing resources of the Dutch grid, which is part of EGI, are accessed using the gLite middleware and the VLEMED virtual organization. The data resources are located at the AMC and managed by the various research groups, due to privacy and intellectual property restrictions. Various custom protocols are used by the data

[9]HELIO Project Page - http://www.helio-vo.eu/

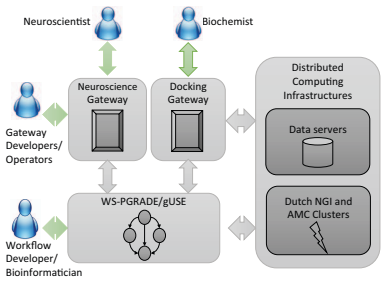

Fig. 3. Infrastructure for large-scale biomedical research at the AMC.

servers. The infrastructure and requirements of this community is characterized by variety. SWMS's are an important stepping stone for this community. They are daily used both directly by biomedical scientists (to build their own experiments) or by science gateway developers (as the backbone of the customized science gateways).

The first portal (2011-2013) was based on MOTEUR[10], and it included applications for processing neuroimaging, sequencing, and mass-spectrometry data [23]. Currently three portals based on gUSE/WS-PGRADE are available for AMC researchers. The generic portal is connected to the Dutch grid and some AMC local clusters and used by advanced users for workflow development and execution. Two customized portals are also available: analysis of neuroimaging data [24] and for virtual drug screening via simulated docking with AutoDock Vina [12]. A detailed analysis of the characteristics of workflows for biomedical applications that have been used in these various portals is presented in [16].

Figure 3 represents an overview of the current infrastructure for biomedical research at the AMC. It illustrates the distributed computing infrastructure, the main systems and the people involved in this eco-system. The neuroscientists and biochemists are researchers who are interested in performing some pre-defined computation on a large data collection. For example, to segment the brain regions from structural Magnetic Resonance Imaging (MRI), which can take more than one day to compute for each scan. Or to calculate the affinity between proteins and ligands, which requires long simulations. These users require a user-friendly interface to upload their data or select it from their local data server, to choose the application to execute on the data, to monitor its execution, and to retrieve the results. The workflow developers use the generic interfaces offered by WS-PGRADE for implementing new workflows (applications). Bioinformaticians also adopt this generic interface because it gives them freedom to develop new data analysis pipelines. The science gateway developers and operators, however, use the SWMS from a completely different perspective. For the developer, programmatic interfaces are necessary to communicate with the SWMS for execution, monitoring and debugging of the workflows. Administrators interact with the SWMS at the system level, taking care of installation, configuration and, most importantly, monitoring system health and troubleshooting when necessary. An administrator view has been developed for the AMC gateways to facilitate troubleshooting, linking information at the application, workflow and infrastructure layers.

IV. Users and Requirements

The three communities presented in Section III have specific and unique characteristics, as illustrated by the ecosystems shown in Figures 1, 2, 3. Regarding the workflow systems used, both Astrophysics and Heliophysics use a combination of Taverna and WS-PGRADE to access a varied computing and data infrastructures. The set-up of the Biomedical community was initially based on MOTEUR, but it is currently based only on WS-PGRADE and a variety of data and computing infrastructures. Also the types of workflows differ within and among the communities: some are data-oriented and others are compute-oriented; some perform long computations (e.g. parameter sweeps) and others perform short data manipulations. The duration, patterns and usage scenarios of the workflows also vary a lot. Some workflows are executed only once (e.g. in Heliophysics experiments), whereas others are repeated for different input data (e.g. biomedical data analysis experiments) or for different parameters (e.g. parameter sweeps for astronomical data visualization). From a workflow management perspective, these communities have diverse requirements. However, these do not differ from what is already known from research and practice of this very active e-science field. For example, a large body of work is devoted to the study of workflow patterns, and how these can be supported by workflow management systems [1]. The study presented here focuses instead on the workflow system *usage* perspective, where it is possible to identify remarkable resemblance among these communities, in spite of all their differences.

When comparing the ecosystem of the three communities, we observed that three main user profiles pop-out: *domain experts* (or scientists), *workflow experts* (or developers), and *science gateway experts* (or developers and operators). Note that these profiles may overlap, and it is at times difficult to draw clear-cut borders between one and the other. Also the same person might take various roles, for example in small communities. Nevertheless it is possible to identify some basic roles and analyse the corresponding usage requirements for the "ideal" SWMS.

Following the methodology suggested in [8] we try to isolate the most relevant information flows to each profile and understand which would be the ideal tools to process them. For analysis of the information needs we use the levels in the Model Driven Architecture: *conceptual*, *abstract* and *concrete*. The conceptual level concerns the information regarding the *scientific domain*, which is typically covered by the science gateway interface. The abstract level concerns the *workflow infrastructure*, which is covered by one or more SWMS's. The concrete level concerns *distributed infrastructure*, which is typically hidden from users of the workflow infrastructure. For information exchange we distinguish between human-system and system-system interfaces. Here again we use the classification from [8]: human-system interfaces include GUI, portals, files and scripts, and system-system interfaces include API and web services. See in Figure 4 an overview of the various actors and their information needs. The user profiles and requirements are detailed below.

A. Domain Expert

Scientists (or *domain experts*) operate at the conceptual level, being most interested in the scientific results that a work-

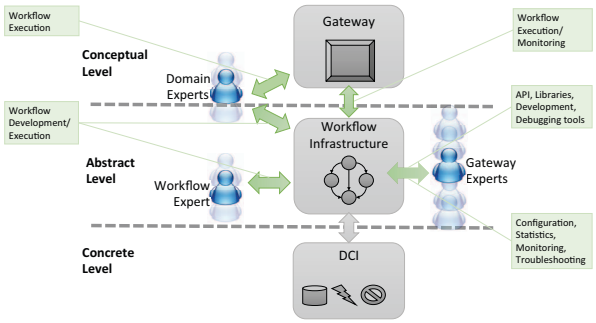

Fig. 4. Ecosystem of users of workflow systems showing information levels (concept, abstract and concrete), main components (gateway, workflow and infrastructure), user profiles (domain, workflow and gateway expert), and main information exchanged at each interface with the workflow system. User profiles can span a wide spectrum of information levels.

flow generates. Additionally, they may have (marginal) roles in the development of workflows, applications and gateways. Prior to the introduction of workflows and portals, domain experts have used scripts and programs. This approach is still actively used in different communities such as Heliophysics where, along Taverna and WS-PGRADE workflows, IDL scripts are used to orchestrate and combine web-services. Apart from scripts and based on our experience, domain experts use two main types of interfaces to science gateways: generic and customized. These are preferably GUIs and portals, but scripting may be also important for more flexibility.

Whenever a customized gateway interface is available, domain experts may access only the information that is filtered and produced in a predefined manner. Typically the user runs an existing workflow on different data or parameter values. Details of the computing or the workflow infrastructure are not relevant for this type of user. Should the execution fail, the user is given only information that is relevant at a conceptual level, upon which he/she user can act; e.g., fail due to wrong data or parameters. On the other hand, details about failure due to the workflow execution engine or a problem in the computing infrastructure are not interesting for the domain scientist.

Domain experts that are also interested in the workflow technology itself (e.g. workflow development and execution) can use the generic and more powerful interfaces offered by the workflow system. This enables the scientist to create new data processing methods and experiments. For example, WS-PGRADE enables modular compositions of meta-workflows from available libraries of sub-workflows (hierarchical workflows). It also enables making specific workflows available for other domain experts with rapid prototyping of simple user interfaces (coined "Applications"). We have observed that some scientists become motivated to learn more about workflow development when the science gateway interface is too restrictive for their needs. This transition is however not straightforward, and in the case of our communities it is typically done in collaboration with a workflow expert.

B. Workflow experts

Workflow experts design and develop workflows that can be later used by themselves or other domain experts. They might overlap with scientists that are interested in technology, or with experts that port applications for science gateways. Workflow experts typically operate at the abstract level; in practice, however, they also need to be aware of the infrastructure and services that will execute the workflows.

Workflow experts are akin to software developers, but constrained to a very specialized environment. They need tools and interfaces to shorten development and debugging time, but they normally have less experience in software development to rely upon. Workflow development is further complicated by the variety of workflow languages and platforms, and by the rise of workflow interoperability platforms such as SHIWA, which make it possible to use them all together. All these systems have their own, typically steep, learning curve. Depending on the language selected, the entire approach to the development can significantly change. As an example, a Taverna developer has a single representation of the workflow where all the aspects of the computation coexist. A WS-PGRADE developer, however, has four different representation of a workflow (abstract, concrete, template and simple user interface) that can be used to define different facets of the problem (e.g., topology of the workflow, implementation details for each node, execution details for each node). A workflow developer needs tools similar to an IDE (Integrated Development Environment) that offer effective user interfaces for various tasks: to browse components of the workflow, either sub-workflows, web services and executables; to compose, navigate and edit the workflow, including tools for zooming into different levels of the workflow; to select editing and deleting single components; to easily modify the graph topology; and to execute and debug the workflows. These users typically prefer GUIs and portals, however scripting and support for repetitive tasks can be useful for heavy users.

C. Science gateway expert

The gateway expert profile is related to the installation, configuration and maintenance of the portal (the *operator*), as well as to the design and implementation of graphical web user interfaces to configure, submit, monitor and access the results of workflows (the *developer*). Here we consider science gateways that is connected to different computing services (e.g. grid, local cluster, cloud) via a SWMS. The gateway experts operate between the abstract and the concrete levels.

Creating a gateway requires domain experts to define the requirements the gateway must meet, workflow experts

to create the workflows, and finally gateway developers to implement the user interface and to link it to the workflow system to run these workflows. The science gateway hides the workflows, therefore the way in which the users interact with the workflow is defined by the gateway developer. Either APIs or web services are used as interfaces between the gateway and the workflow system. These interfaces make workflow submission, monitoring and collection of the results very easy. These interfaces should allow rich information exchange for control and monitoring purposes, for example, with detailed status and progress information about all tasks executed by the workflow enactment in the DCI. The gateway expert should design the front-end as a robust and maintainable system, based on modular and reusable design. For example gateway experts may take advantage of the modularity of Liferay to create reusable portlets and employ the Application Specific Module (ASM) API of WS-PGRADE/gUSE to manage the workflows. Since gateways are very complex systems, the SWMS should also offer possibilities of debugging and troubleshooting the communication between the various components. Note that an extra layer on top of the workflow system may be still needed to capture the provenance of data and experiments from the conceptual perspective, because the data infrastructure and its semantics are typically not known to the workflow system.

The gateway operator takes care of installing, monitoring the services and troubleshooting, being also responsible for the service availability according to the QoS defined with the users. From the workflow system perspective, the operator needs interfaces to facilitate DCI configuration, identification and solution of the problems encountered during workflow execution. In practice this can be challenging because the abstract layer introduced by the workflow infrastructure may hamper debugging and troubleshooting, as it may hide too much concrete level information. The operator also requires interfaces to activity information, for example, statistics about usage and failures. GUI, portals and scripts are possible interfaces between the operator and the SWMS.

V. Discussion

The experiences gathered so far by the scientific communities, described in Section III, tell about the challenges and solutions faces in the daily management and development of a workflow-centred e-Science infrastructure. It is interesting to observe how communities that different wildly in size, technology and usage patterns had to face the same problem of isolating and managing information at the right level in the various interfaces between the workflow system and the external human and software components.

As summarized in Figure 4, a large variety of actors and information needs come together in a workflow-based e-science environment or science gateway. Each one of the actors, however, has specific information needs. For example, the domain experts using the provided interfaces are interested only on the success or not of the execution of an experiment; he/she must be bothered solely about inputting data and retrieving results. To the scientist, a failure is relevant only if there is an error in the input datasets, which has meaning in the scientific context. On the other end, for a gateway expert who maintains a science gateway, the detailed reason for a failure of a workflow is very important, to enable identification and solution of the problem, as well as to establish credibility to the results. These two contrasting requirements - to hide and to expose details - are challenging. They need to be addressed with different mechanisms for information exchange at the conceptual, abstract and concrete levels between the workflow infrastructure and its various types of "users."

Although there has been no attempt to formalize a common methodology, it is interesting to observe how these three communities have addressed these challenges with similar approaches. For example, the Heliophysics and Astrophysics communities developed Taverna plugins that offer common services for the orchestration of the web services. These plugins facilitate workflow development by enriching the workflow system with concepts from these application areas. Moreover, they independently developed an approach to reuse workflows based on the concepts of meta-workflows and workflow interoperability, whereby high-level scientific workflows are built by the composition of specialized sub-workflows. The atomic or sub-workflows implement a simple task that can be represented by a simple application; they hide all the technical details within the workflow and propagate only results and meaningful exceptions. A similar approach also starts to be used by the Biomedical research community, to exploit existing web services for enrichment of genomics data in combination with large capacity offered via WS-PGRADE. These communities recognized the need to leave freedom of choice of SWMS to their users, and adopted a workflow-interoperability approach. Finally, the communities implement both customized portlets for a particular user group, or use the different views of WS-PGRADE (e.g. *End User View*) to offer constrained interfaces that manage information relevant to the domain experts. The Biomedical community has also implemented a specific interface for gateway administration, where information of various workflow levels (conceptual, abstract and concrete) are linked to facilitate troubleshooting.

We are aware that in this paper we have only studied the observations from three user communities, which is quite a limited sample. However the approach followed "naturally" by these communities is actually not limited to their experiences. In the literature it is possible to identify comparable examples. The construction of science gateways based on SWMS is common practice in various scientific domains (see list in [14]). For example, MoSGrid, a computational chemistry community that also participates in ER-flow and SCI-bus projects, has been heavy user of workflows for a long time, and have developed a sophisticated and successful science gateway. Their organization, with developers and scientists, is similar to ours. Moreover, they are also developing atomic workflows as a toolbox to enable scientists to more easily develop their own meta-workflows at the conceptual level [11]. On the other hand, also in the Teragrid and XSEDE science gateway initiatives, the infrastructure providers identify the roles of scientists users and science gateway experts [18],[27]. The approach identified in this paper, where three roles are identified for users of SWMS's, seems to be natural and possibly extends to different communities that have faced the same problems with the similar technological approach.

VI. Conclusions

The experience gathered so far in the collaborations within the ER-Flow and SCI-BUS projects has highlighted an interesting pattern across the different communities. Faced with similar challenges, they independently adopted similar solutions trying to isolate, manage and abstract information flows to gather the needs of different user profiles (domain, workflow and gateway experts). We have shown that these diverse user profiles have demands that go beyond the requirements expressed so far for the information exchange between a human user ("the scientist") and the SWMS. The need for adequate system-system communication at various information levels indicates new requirements for SWMS.

Each community has naturally followed a design and implementation approach that tried to isolate different layers in each system and to offer optimized interfaces to each user profile. This approach, albeit not formalized, resembles the *Separation of Concerns* design principles, and aspect-oriented programming in particular. These design principles have been adopted in the model formalized in [8], which proposes implementing workflows with three levels of abstraction to facilitate the isolation and management of the different information flows that are woven within any complex worfkflow-based eScience infrastructure. It would be greatly beneficial if the practical experience gathered by these workflow user communities and the formal approaches proposed so far could be united for the design of novel workflows infrastructures to effectively support its various user profiles in a more effective way. Given the increasing interest in the construction of science gateways that are powered by workflow management infrastructure, it is expected that such benefit could impact a large number of communities and e-science environments.

Acknowledgments

The authors thank colleagues from the INAF IT group who participated to the development and design of VisIVO workflows and science gateway, in particular F. Vitello, P. Massimino and A. Costa. The authors are grateful to the HELIO team for their support, to J. Walsh for his help in maintaining the infrastructure and to J. Byrne, D. Perez Suarez and P. Higgins who helped with the Science Cases. The authors also thank the colleagues from the AMC e-science group and collaborators, in particular S. Shahand, M. Santcroos, M. Caan, A. van Kampen and B. van Schaik. The Dutch e-Science Grid is financially supported by Netherlands Organisation for Scientific Research, NWO, and by Stichting SURF. This work was developed within ER-Flow (FP7 INFRASTRUCTURES-2012-1 contract 312579) and SCI-BUS (FP7-INFRASTRUCTURES-2011 contract 283481) projects.

References

[1] Van Der Aalst et al. Workflow patterns. *Distrib. Parallel Databases*, 14(1):5–51, July 2003.

[2] A. Barker and J. Hemert. Scientific workflow: A survey and research directions. In *Parallel Processing and Applied Mathematics*, volume 4967 of *Lecture Notes in Computer Science*, pages 746–753. Springer Berlin Heidelberg, 2008.

[3] U. Becciani et al. Science gateway technologies for the astrophysics community. *Concurrency and Comp: Practice and Experience*, 2014.

[4] U. Becciani et al. Creating gateway alliances using WS-PGRADE/gUSE. In *Science gateways for distributed computing infrastructures*. Springer, in press.

[5] R. Bentley et al. HELIO: Discovery and analysis of data in heliophysics. *Future Generation Computer Systems*, 29(8):2157–2168, 2013.

[6] A. Blanc et al. Workflows for heliophysics. *Journal of Grid Computing*, 11(3):481–503, 2013.

[7] N. Cerezo, J. Montagnat, and M. Blay-Fornarino. Computer-assisted scientific workflow design. *J of Grid Computing*, 11(3):585–612, 2013.

[8] Nadia Cerezo. *Workflows conceptuels*. PhD thesis, University of Nice Sophia Antipolis, 2013.

[9] E. Deelman, D. Gannon, M. S. Shields, and I. Taylor. Workflows and e-science: An overview of workflow system features and capabilities. *Future Generation Computer Systems*, 25(5):528–540, 2009.

[10] T. Glatard, J. Montagnat, D. Lingrand, and X. Pennec. Flexible and efficient workflow deployement of data-intensive applications on grids with MOTEUR. *International Journal of High Performance Computing Applications*, 22(3):347–360, 2008.

[11] S. Herres-Pawlis et al. User-friendly workflows in quantum chemistry. In *Proc. Int. Workshop on Scientific Gateways (IWSG)*, 2013.

[12] M. Jaghoori et al. A grid-enabled virtual screening gateway. In *Proc. 6th Int. Workshop on Science Gateways (IWSG)*, 2014.

[13] P. Kacsuk et al. WS-PGRADE: Supporting parameter sweep applications in workflows. In *Proc. 3rd Workshop on Workflows in Support of Large-Scale Science. (WORKS)*, pages 1–10. IEEE, 2008.

[14] H. Nguyen and D. Abramson. Workways: Interactive workflow-based science gateways. In *IEEE 8th International Conference on e-Science*, pages 1–8, Oct 2012.

[15] T. Oinn et al. Taverna: Lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice & Experience*, 18(10):1067 – 1100, 2006.

[16] S. D. Olabarriaga et al. Understanding workflows for distributed computing: Nitty-gritty details. In *Proc. of the 8th Workshop on Workflows in Support of Large-Scale Science*, WORKS, pages 68–76, New York, NY, USA, 2013. ACM.

[17] G. Pierantoni, B. Coghlan, and E. Kenny. The architecture of HELIO. In *Krakow Grid Workshop*, 2010.

[18] M. Pierce et al. Open grid computing environments: Advanced gateway support activities. In *Proceedings of the TeraGrid Conference*, pages 16:1–16:9, New York, NY, USA, 2010. ACM.

[19] Planck Collaboration, Ade, P. A. R., Aghanim, N., Arnaud, M., et al. Planck early results. i. the planck mission. *A&A*, 536:A1, 2011.

[20] A. Schaaff et al. Scientific Workflows in Astronomy. In *Astronomical Data Analysis Software and Systems XXI*, volume 461 of *Astronomical Society of the Pacific Conference Series*, page 875, September 2012.

[21] E. Sciacca et al. Visivo science gateway: a collaborative environment for the astrophysics community. In *5th International Workshop on Science Gateways, IWSG 2013*. CEUR Workshop Proceedings, 2013.

[22] E. Sciacca et al. Visivo workflow-oriented science gateway for astrophysical visualization. In *Parallel, Distributed and Network-Based Processing (PDP), 2013 21st Euromicro International Conference on*, pages 164–171. IEEE, 2013.

[23] S. Shahand et al. A grid-enabled gateway for biomedical data analysis. *J. of Grid Computing*, 10(4):725–742, 2012.

[24] S. Shahand et al. A data-centric neuroscience gateway: design, implementation, and experiences. *Concurrency and Computation: Practice and Experience*, 2014.

[25] G. Taffoni et al. Enabling grid technologies for planck space mission. *Future Gener. Comput. Syst.*, 23(2):189–200, February 2007.

[26] F. Vitello et al. Developing a mobile application connected to a science gateway. In *6th Int. Workshop on Science Gateways, IWSG*, 2014.

[27] N. Wilkins-Diehr et al. Teragrid science gateways and their impact on science. *Computer*, 41(11):32–41, November 2008.

[28] Workflow Management Coallition. Workflow Reference Model, 1995.

[29] Gil Y. et al. Examining the challenges of scientific workflows. *Computer*, 4(12):24–32, 2007.

[30] J. Yu and R. Buyya. A taxonomy of workflow management systems for grid computing. *Journal of Grid Computing*, 3(3-4):171–200, 2005.