Project Number: **RI-312579**
Project Acronym: **ER-flow**

Project Full Title:
Building an European Research Community through Interoperable Workflows and Data

Theme: Research Infrastructures
Call Identifier: FP7-Infrastructures-2012-1
Funding Scheme: Coordination and Support Action

Deliverable D*5.5*
Description of applications ported to the SSP (year 2)

Due date of deliverable: 28/08/2014 Actual submission date: 31/10/2014
Start date of project: 01/09/2012 Duration: 26 months

Lead Contractor: University of Westminster
Dissemination Level: PU
Version: 1.0

ER-flow is supported by the FP7 Capacities Programme under contract no RI-312579

1 Table of Contents

Contents

2 List of Figures

3 List of Tables

4 Status and Change History

Status:	Name:	Date:	Signature:
Draft:	Gabriele Pierantoni	16/08/2014	n.n. electronically
Reviewed:	Silvia D. Olabarriaga	28/10/2014	n.n. electronically
Approved:	Gabor Terstyanszky	30/10/2014	n.n. electronically

Table 1, Deliverable Status

Version	Date	Pages	Author	Modification
0.1	26/11/13	All	GP	Created Skeleton
0.2	16/08/14	All	GP	Draft
1.0	29/09/14	All	GP, all	Final Version – Structure
1.1	30/09/14	All	GP	Final Version – Errors corrected
1.2	16/10/14	All	GP	Final Version – Improved version to meet comments from the reviewers
1.3	21/10/14	All	GP, all	Final Version – Astro Section final, Computational Chemistry Final, Porting Experience.
1.4	27/10/14	All	GP, all	Gabor's comments addressed by all the communities

Table 2, Deliverable Change History

5 Glossary

CIR	co-rotating interaction region
CME	Coronal Mass Ejections
CMB	Cosmic Microwave Background
DCI	Distributed Computing Infrastructures
DNA	Deoxyribonucleic acid
DTI	Diffusion Tensor Imaging
ETA	Expected Time of Arrival
GR	General Relativity
gUSE	grid User Support Environment
HEC	Heliophysics Event Catalogue
HFC	Heliophysics Feature Catalogue
MoSGrid	Molecular Simulation Grid
MRI	Magnetic Resonance Imaging
MoU	Memorandum of Understanding
NGS	Next generation sequencing
NBS	Nicolaides-Baraitser syndrome
REST	Representational state transfer
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
SMARCA	SWI/SNF-related, Matrix-associated, Actin-dependent Regulator Chromatin
SSP	SHIWA Simulation Platform
VO Table	Virtual Observatory Table

WP	Work package
XML	Extensible Markup Language
XNAT	Extensible Neuroimaging Archive Toolkit

Table 3, Glossary

6 Abstract

This deliverable describes the applications ported to the SHIWA platform during the 2nd year of activities of Work Package 5. This constitutes the main technical activity of the ER-FLOW project. The deliverable covers both technical and scientific details. It contains background about the scientific domains addressed by the applications; the relevance of the selected applications in their respective fields through the description of science cases; the technical characteristics of these applications and the distributed computing infrastructures where they run; general explanations about the porting of these applications as workflows to the SHIWA platform; and how these ported applications are used in the various execution environments. Additional technical details about the workflows are included as appendix.

7 Introduction

The FP7 "**Building a European Research Community through Interoperable Workflows and Data**" (ER-flow) project disseminates the achievements of the SHIWA project[1] and uses these achievements to build workflow user communities across Europe. ER-flow provides application support to research communities within and beyond the project consortium to develop, share and run workflows with the SHIWA Simulation Platform (SSP).

One important work package of ER-FLOW is WP5, the Application Support work package. It deals with the creation and porting of applications of four different communities to the SHIWA Simulation Platform. This deliverable (D5.5) is issued at the end of the second project year and aims at describing the porting process in this period.

The report provides three different levels and types of information about the application support:

- For **domain scientists** interested in an overview of the ported workflows and in the general concepts on their designs, we recommend reading the execution environment and usage patterns sections of this deliverable.
- For **domain scientists** interested in the scientific approach to a particular problem, and in a generic overview of its implementations through workflows, we recommend in particular reading the relevant Science Case sections (The definition of Science Case is detailed later in this section) that are available on the ER-flow web site (http://www.erflow.eu/science-cases)
- For **domain scientists** and **workflow developers** that want more technical details on the workflows and their implementations, we recommend reading the appendix and the Application Templates available in the workflow repository and on the ER-flow website (http://www.erflow.eu/applications).

We have made an effort to use consistent terminology in the project and throughout the document as follows:

- **Application**: some software/code that performs a relevant task in the scientific domain. An application may be implemented as a workflow, sub-workflow or a meta-workflow.
- **Use Case**: term from software engineering used to identify a list of steps defining interactions between a user and a system to achieve a goal.
- **Meta-workflow**: a workflow that combines and invokes at least one sub-workflow
- **Sub-workflow**: a workflow that is invoked by meta-workflows

Due to the large variety of usage scenarios of the different communities, these definitions can be interpreted with a certain degree of flexibility. We have additionally introduced the concept of **Science Cases**, which describe examples of usage of workflows in a specific scientific context. The Science Cases illustrate the scientific relevance, usability and usefulness of the developed/ported workflows. We have defined the Science Case as an extension to the concept of the Use Case, which is widely used in Computer Science. Each community has developed 3-4 Science Cases with the available workflows and has described such Science Cases in the related sections of this deliverable. The variety of science cases provide an indication of the heterogeneity of the problems that can be and have been addressed with the help of scientific workflow management systems in this project. Note that the science cases may refer to workflows developed in the first or the second year, or on a combination of two or more workflows.

[1] http://www.shiwa-workflow.eu/project

The porting process of applications as workflows that can run on distributed computing infrastructures (DCI) realized in WP5 entails several successful sub-tasks:

- To select and understand which applications to port, a decision to be taken respecting different criteria such as maximizing the impact of the ported application and maximizing the added value of the workflow interoperability technology of SHIWA.
- To assess whether the application is still active or is out of date, and, when the application needs updates or changes, also to update them prior to the porting process.
- To arrange all workflows in a fashion to foster re-usability and usefulness. Depending on the different communities this has lead to different solutions ranging from hiding heavily used workflows behind customized web interfaces or the development of meta-workflows.

The porting process itself entails development or adaption of workflow(s) that implement the application, their documentation and publication in the SHIWA Repository, and their deployment for execution from one of the various SHIWA and/or customized execution environments.

This deliverable aims at answering, for each community, the questions highlighted by the previous tasks:

- **Why these applications?** A brief introduction for each of the communities describes the rationale behind the choice of the selected applications and the context in which they are executed.
- **What are these applications?** A concise description of each application is present in the sections dedicated to each of the four communities. More technical details are available in on-line materials that are linked from this deliverable and from the workflow descriptions in the SHIWA Repository
- **How were these applications ported?** A concise report of the process, the experience and challenges related to the porting process are also presented. In fact deliverable D5.1 and D5.3 extensively document this experience, so here we only emphasize aspects that are relevant for each of the specific applications and communities.

This deliverable has been structured with the aim of facilitating direct access to information at different levels of details in the document, which are:

- Document-wide: information that pertains to the entire document:
  - ***Section 7:*** context information of this deliverable.
  - ***Section 12:*** summarizes the porting experience of all the communities
  - ***Section 13:*** presents general conclusions on the porting applications under WP5 or ER-FLOW in the second year of activity.
- Community-wide: information that applies to an entire community is available in the individual sections for each community (**Sections 8-11**):
  - ***Sections x.1:*** generic description of the scientific community, which is adapted/updated from D5.2.
  - ***Sections x.2:*** describe common characteristics of the technical background of the applications of the given community, such as middleware, workflow languages and systems, etc. This section is adapted/updated from D5.2.
  - ***Sections x.3:*** describe common features on how applications are used in the community.
  - ***Sections x.4:*** describe the scientific investigations performed with the workflows developed during years 1 and 2.
    - ***Sections x.4.1:*** brief description of the Science Case.

        - ***Sections x.4.2:*** scientific relevance of the Science Case for the community.
        - ***Sections x.4.3:*** steps that implement the Science Case.
        - ***Sections x.4.4:*** one example of concrete scientific study to illustrate the Science Case.
        - ***Sections x.4.5:*** list of publications related to this Science Case, at the technical and domain levels.
        - ***Sections x.4.6:*** contact persons for more information about the Science Case.
    - ***Sections x.5:*** summarize in a table the workflows created and ported in Year 2. The technical details can be found in four annexes at the end of this deliverable, one for each community.
- Application-wide: each application/workflow is described in detail in four different annexes, one for each community (Annex A – Astrophysics, Annex B – Computational Chemistry, Annex C – Heliophysics, Annex D – Life Sciences). These annexes are structured as follows:
    - ***Nature and Relevance:*** describes features and relevance of the ported application in the scientific domain.
    - ***Workflow Details:*** details of the workflow.
    - ***Software Details:*** details of the software used by the application.
    - ***Input/Output:*** describes data that us used and generated by the workflow
    - ***Further Technical Details:*** covers technical details of interest to potential workflow developers that want to reuse the workflow. These may be subject to change to reflect upgrades in the infrastructure or the application. This section links to the workflows in the SHIWA repository, and other relevant information for workflow reuse.

Note that application templates, which provide additional details about each of the applications, are available on the ER-flow web site (http://www.erflow.eu/applications), as well as linked from the respective workflow(s) in the repository.

8 Astrophysics

Astronomy is a natural science that deals with the study of celestial objects (such as moons, planets, stars, nebulae, and galaxies); the physics, chemistry, mathematics, and evolution of such objects and the phenomena that originate outside the atmosphere of Earth (such as supernovae explosions, gamma ray bursts, and cosmic background radiation).

Astrophysics is the branch of astronomy that deals with the physics of the universe, including the physical properties of celestial objects, as well as their interactions and behavior. The studied objects include galaxies, stars, planets, extra-solar planets, the interstellar medium and the cosmic microwave background. Cosmology is also a branch of Astrophysics that deals with the study of the origins and evolution of the Universe. Astrophysics has become a data intensive science due to numerous digital sky surveys, with many terabytes 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 petabytes of data to be stored and analyzed.

Handling and exploring these vast amounts of 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. Grid, cloud and data e-infrastructures provide a vital foundation for the Astrophysics community. Examples are the European Grid Infrastructure and the Open Science Grid[2], for computing resources, and the Virtual Observatory (IVO) data infrastructure of the International Virtual Observatory Alliance[3] (IVOA), for tools, software and services to access, share, manipulate and visualize data. Workflows systems have been widely used to coordinate services and to access computing resource and data storage resources. Workflows for Astronomers and Astrophysicists are a mean to design applications, access data using IVOA standards, develop science gateways that simplify the access to DCIs and Cloud infrastructures.

To demonstrate the benefit in using workflows systems we identified three science cases dealing with different aspects of the research in A&A. The use cases and associated workflows were setup during Y1 and evolved during the project implementing the new features offered by the SHIWA platform. In parallel, various lightweight data oriented workflow modules have been developed to access tools and services of the IVOA giving access to data resources.

8.1 Technical Background

The proposed scientific use cases cover different A&A disciplines: Astronomy, Astrophysics, Cosmology, Stellar evolution and Astroparticle physics applications. These applications behave differently also from the computational point of view: some applications perform numerical simulations where input data are much smaller than output data, other run data reduction and visualization (input data are much larger than output data) some others are semi-analytical codes where large number of runs of the same code with different input parameters is executed.

All applications are designed as workflows and proposed to the users through science gateways. They were developed using the SHIWA Simulation Platform and the gUSE/WS-PGRADE technology.
The distributed infrastructure used to run the applications is the EGI Grid Infrastructure. The middleware used is the gLite middleware as released by EMI. The A&A community operates

[2] www.opensciencegrid.org
[3] http://www.ivoa.net

three different virtual organizations: planck (set up to fulfil the computing and storage needs of the Planck satellite); inaf (the Italian National Institute of Astrophysics VO) and the astro.vo.eu-egee.org (generic VO maintained by EGI in collaboration with the French NGI). The overall amount of sites involved in the A&A VOs is 50 distributed all over Europe.

The workflows and portals have been setup to access these three VOs and the A&A WS-PGRADE workflows, to coordinate the Grid computing elements and Storage elements and to allow to access and store files in the Grid distributed file catalogue (LFC).

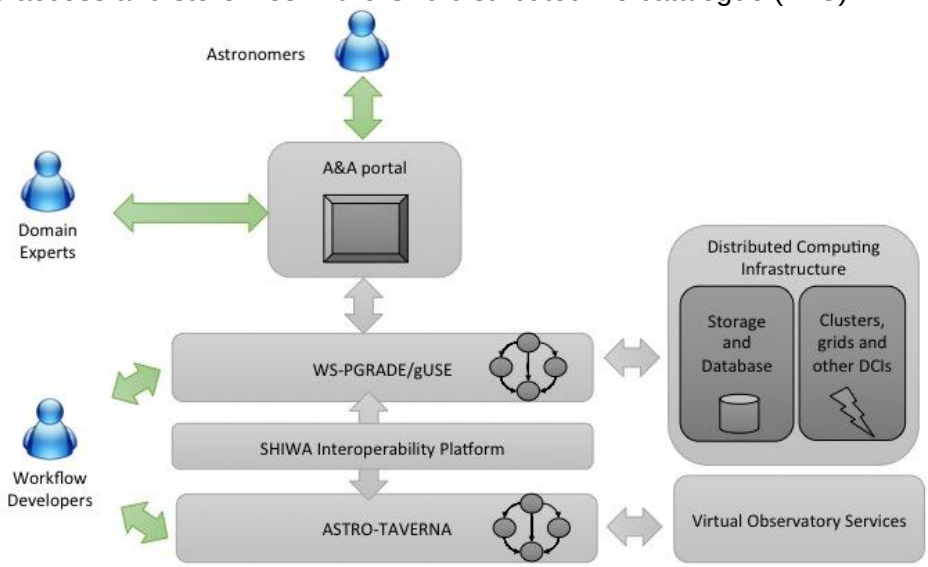


Figure 1, Execution environment of A&A workflows

While the astro workflows described above are strictly computing oriented, it is mandatory for Astronomers to access data using IVOA standards and services. The European Virtual Observatory[4] (Euro-VO) provides access to A&A data centres thanks to a set of tools and services based on REST web services technology. These services have been developed in the framework of different EU funded projects such as Euro-VO CoSADIE[5]. During Y2 we also used cloud and local resources as workflow back-ends. For example the STARnet nodes use cloud and local resources.

To develop workflows that also allow data access we use the SHIWA Simulation Platform capability to execute non-native workflows. The simulation platform also enables the development and execution of non-native workflows, i.e. workflows built through workflow systems other than WS-PGRADE, e.g. the TAVERNA[6] workflows using the AstroTAVERNA[7] plugin. The TAVERNA workflows are of particular interest for the Astrophysics community, as tens of workflows already exist and are used inside the community, given that they allow an easy use and exploitation of the features provided by the Virtual Observatory[8]). We tried AstroTAVERNA workflows given their ability to manipulate the VOTable format proving that they work properly. All these non-native workflows do not access directly the middleware or the computational and storage resources or any other grid resource; therefore there is no need for grid certificates or affiliation to any Virtual Organization. The resources used by the workflows are reached through the AstroTAVERNA plugin, and the workflows runs take place on a local cluster at the University of Westminster.

[4] http://www.euro-vo.org
[5] http://www.cosadie.eu
[6] http://www.taverna.org.uk
[7] http://amiga.iaa.es/p/290-astroTAVERNA.htm
[8] http://www.ivoa.net

During the Y2 of the project a federation of Astrophysics-oriented science gateways, named STARnet, has been designed and implemented. STARnet is based on gUSE/WS-PGRADE gateway platform. STARnet aims sharing a set of services for authentication (based on OpenLdap, OpenID, SAML) and a common and distributed computing infrastructure (clusters or DCIs), data archives and workflow repositories (based on OwnCloud with STARnet data replication and synchronization service). The first implementation of STARnet involves 5 gateways:

- INAF - Osservatorio Astrofisico, Catania (OACT) Italy[9].
- INAF - Osservatorio Astronomico, Teramo (OATE) Italy[10] .
- INAF - Osservatorio Astronomico, Trieste (OATS) Italy[11].
- University of Portsmouth, UK (UP)[12]
- Astronomical Institute of Slovak Academy of Sciences (AISAS)[13] .

8.2 Workflow Usage in the Community

Participating in the ER-flow and SCI-BUS projects, the Astrophysics community has gained experience in workflow design and implementation as well as usage of science gateways. The A&A community uses three different kinds of workflows: *data-oriented* workflows, *visualization-oriented* workflows and *computing-oriented* workflows.

- The **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. They use IVOA standards and REST web services. These lightweight operations are normally implemented by an individual simple workflow module that does not access directly DCIs or any (grid) computational and storage resources; these modules are simple to operate and not computing intensive. These workflows are highly re-usable as they implement operations that can be embedded in a number of applications/services that require data access.
- **Visualization-oriented workflows** are used to explore and visualize large data sets.
- **Computing-oriented workflows** consist of computing tasks that need to be organised into a complex workflow. In this case we identified two different workflow patterns. The first pattern consists in running multiple instances of the same workflow on different inputs, exploring different parameters. The second pattern consists in analysing different data using the same workflow, which is the case of data reduction/analysis pipelines.

There are two types of users in the A&A community represented in ER-flow: advanced users that develop and run workflows (workflow developers) and standard users that execute workflows from a customized web interface (domain scientists). The first type of users access WS-PGRADE portals to design and implement the workflows. The standard users access A&A customized science gateways (the STARnet gateways) where an application environment has been developed using Java portlet technology.

Workflow developers design and develop workflows that later can be used by themselves or by other scientists. They are akin to software developers, but constrained to a very specialized environment. The science gateways are installed, maintained and configured by a science gateway expert that is also in charge of the design and implementation of the graphical web user interfaces.

[9] http://visivo.oact.inaf.it:8080/
[10] http://193.204.1.135:8081/
[11] http://guse-fe.oats.inaf.it:8080/
[12] http://148.197.12.1:8081/
[13] http://sg-mph.ta3.sk:8081/

8.3 Science Cases

We have considered the most interesting science cases belonging to the first year of ER-flow and developed in WS-PGRADE; the workflows of the second year are more basic, developed with a different technology (TAVERNA + AstroTAVERNA plugin), and thought for future science cases.

The science cases developed by the Astrophysics community in the ER-flow project are summarized below.

Name	Description	Workflow IDs
Virtual Observatory	Virtual Observatory Interface	See Table 5
Planck	Simulations of the ESA Planck satellite mission	4978
Visualization	Muon workflow and gateway for displaying tomographic images	5103

Table 4, Astrophysics Science Cases

8.3.1 Virtual Observatory use case

The Virtual Observatory produces a unified virtual data and service resource with the ability to perform complex data discovery and manipulation of tasks across the whole range of astronomical resources provided by distributed Astronomical Data Centres. The IVOA developed a set of standards, services and applications that are commonly and widely used by Astronomers to access archives and catalogues and to perform operations on observed data.

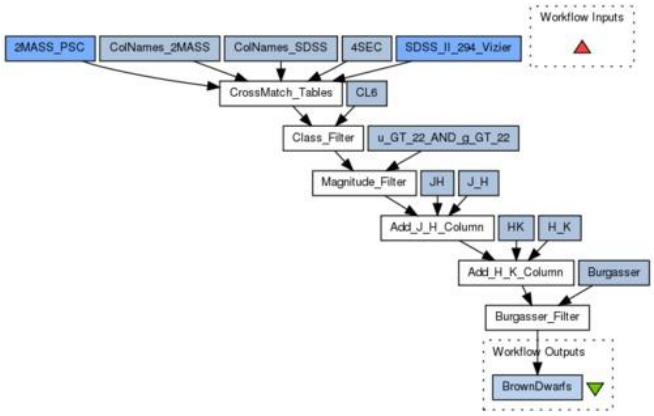


Figure 2, Discovery of brown dwarfs mining the 2MASS and SDSS databases

To develop science gateways that enable Astronomers to access at the same time computing and IVO data resources, it is necessary to implement a set of data-oriented workflows to access Virtual Observatory resources. We implemented workflows based on AstroTAVERNA plugins as data-oriented workflow modules that are used to access and manipulate data.

In particular, we implemented both AstroTAVERNA Plugin workflows as simple atomic operations that can be re-used as basic components (sub-workflows) of other more complex meta-workflows that represent Astronomical use cases. Moreover those atomic workflows are used also in the implementation of SSP science gateways as simple operations.

Workflow	Type
Discovery of Brown Dwarfs mining the 2MASS and SDSS databases	Application
Retrieving information from HST ConeSearch and Image VO Services	Application
Concatenates several VOTables into one	ATOMIC
Create configuration files from a template and a VOTable	ATOMIC
Astronomical object name to equatorial coordinates Resolver	ATOMIC
Run scripts from a column in a VOTable	ATOMIC
Create VOTable from ellipse results	ATOMIC
Add columns to a VOTable resulting from the execution of sextractor	ATOMIC
Extract a column from a VOTable into a List	ATOMIC
Perform a ConeSearch query to a VO Service	ATOMIC
Split VOTables into its values	ATOMIC
Extract content of columns from VOTables	ATOMIC
Find events in x-ray and radio	Application
Create a VOTable of NED (NASA/IPAC Extragalactic Database) images from a list of objects	ATOMIC
Galaxies Sample Selection Research Object	Application

Table 5, IVO workflow name and type

8.3.1.1 Scientific Merit

Making the DCI and IVO resource interoperable is a key activity for astronomers since the introduction of the grid e-infrastructures in Europe. Those workflows help in solving this problem as they give SSP the ability not only to interact with DCI resources but also with IVO ones, hiding the complexity of the computing and data infrastructure and allowing to focus only on the scientific problem. Moreover it gives to Astronomical Data Centers the unique opportunity to develop SSP based science gateways to access data using IVO standards/services and to analyze and reduce these data on DCIs.

During Y2 we built a “library” of “atomic” workflows based on AstroTAVERNA to use as basic modules of more complex workflows. We implemented a few AstroTAVERNA based workflowsas shown in Table 5; we discuss here the steps of one application (HST cone search) and one atomic workflow (VOTable Concatenation).

8.3.1.2 Steps

The Hubble Space Telescope (HST) archive offers access to the HST observations using a VO service. Data can be retrieved in the form of FITS files or VOTables. The cone search workflow is presented in Figure 3. This workflow performs a cone-search around a point in the Sky. The input of the workflow is an ASCII file with a list of source names. First this file is converted into a VOTable then an extra-column is added for the coordinates that are necessary to query the SIAP image service. The HST SIAP service is queried. The final result consists in two different VOTables issued as the response of both VO Services; Images VOTable has been previously filtered.

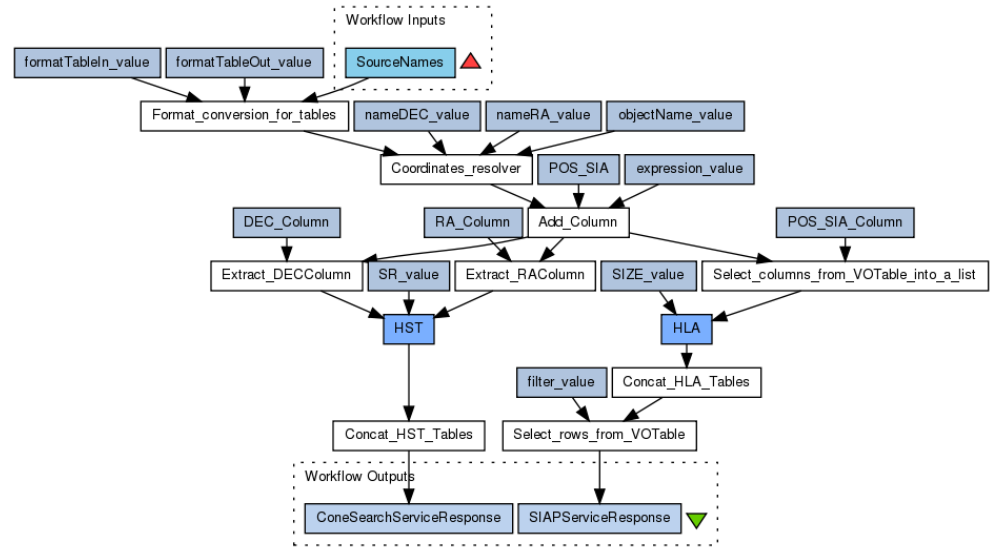


Figure 3, Retrieving information from HST Cone Search workflow

One of the common tasks to do when working on astronomical data is to combine data retrieved from archives or catalogues. The workflow presented in Figure 4 is an example of atomic workflow that can be implemented as a light-weight module into more complex workflows. It accepts VOTables as input data then it combines them into a vertically replicated singe table.

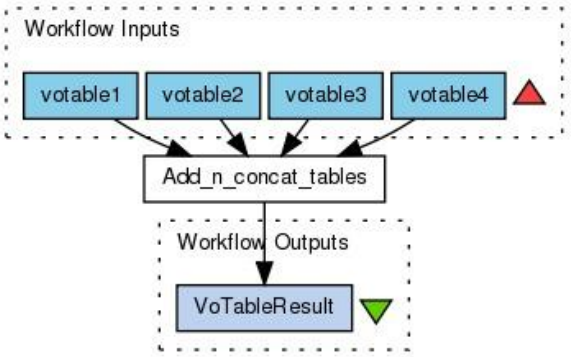


Figure 4, Concatenate VOTables

8.3.1.3 Example

We queried the HST cone search service from Space Telescope for records within .05 degrees of each Messier object contained in a local input file. The sky positions in the input catalogue are guessed from the available table metadata.

The result is a catalog of objects with their coordinates (RA, DEC), the instrument used, the date and characteristic of the observation. The link to the associated images and image type (DARK, BIAS, FLAT, IMAGE) is also provided. We calibrated the images and we made the photometric analysis. In Figure 5 we present the result of a calibrated image at RA: 03 32 21.90 and Dec: -27 49 29.82 observed the 4th of September 2005 at 9:40PM; the calibrated image has been queried from the HST archive. The exposure time is 1000 seconds, the instrument used is the Wide-Field Planetary Camera 2 and the filter adopted is the F300W.

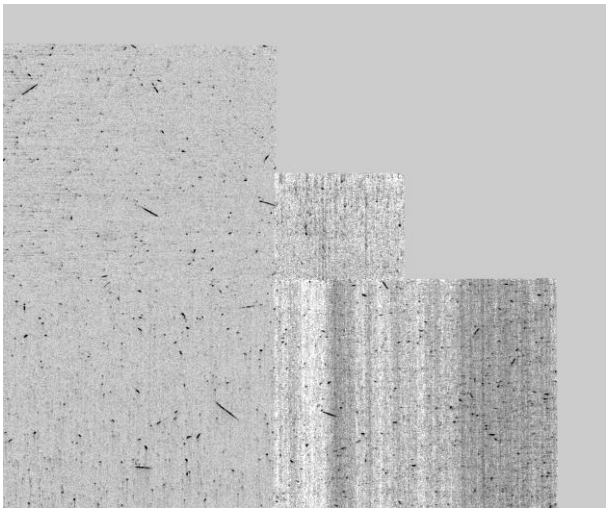


Figure 5, HST WSPC2 calibrated image

8.3.1.4 Related Publications

Castelli, G., Taffoni, G., Sciacca, E., et al. “VO-compliant workflows and science gateways”, 2014, A&C submitted.

8.3.1.5 Contacts

- Giuliano Taffoni: taffoni@oats.inaf.it
- Riccardo Smareglia: smareglia@oats.inaf.it
- Eva Sciacca: sciacca@oact.inaf.it

8.3.2 Planck Science Case

This application relates to Astronomy, Astrophysics and Cosmology, and it is an evolution of the year 1 Planck workflow that has been re-implemented in year 2 as a meta-workflow. The application runs the simulations of the Planck LFI mission. The computing and storage needs related to the simulations of the Planck mission imply the use of Grid DCIs. In particular we used the capability of the SSP to access gLite resources (Planck Virtual Organization). Moreover the application must be executed a large number of times varying the input parameters.

To explore different parameters a set of Planck simulations, workflows are executed in parallel on different resources, and thereafter the outputs are combined. The output data is produced in VOTable format and then combined into a single VOTable. The VOTable format is used in order to obtain IVOA compliant data. The meta-workflow implements native WS-PGRADE workflows and AstroTAVERNA ones as shown in Figure 6.

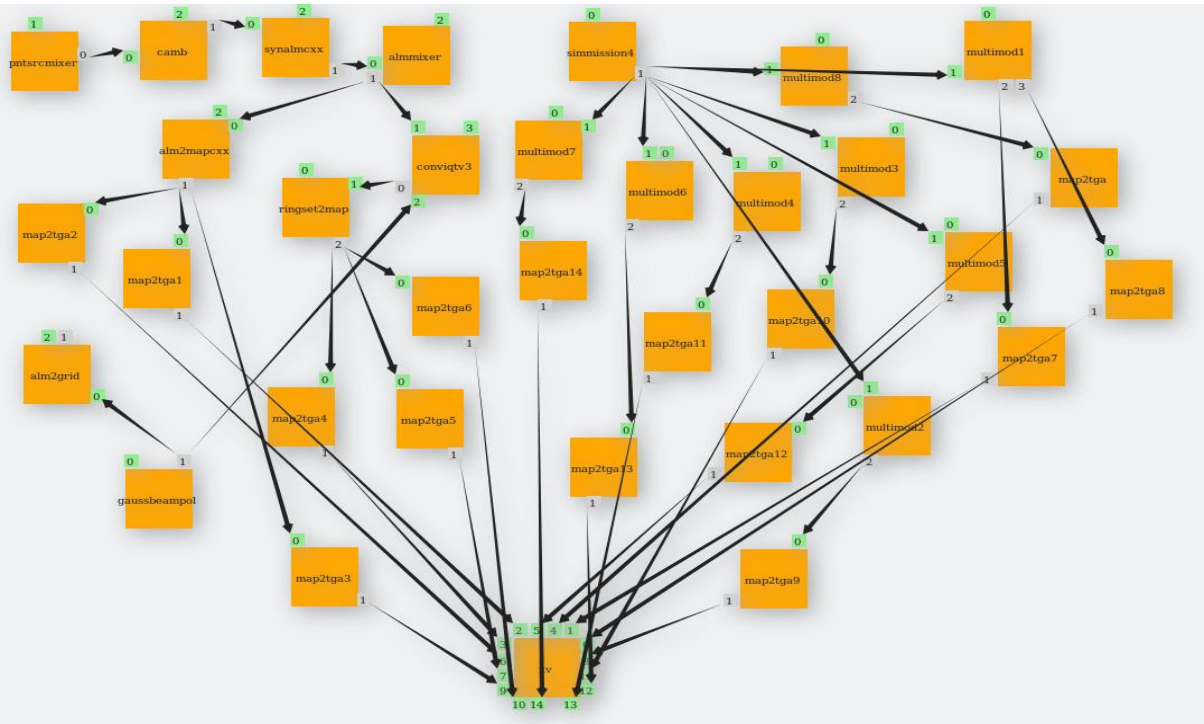


Figure 6, PLANCK Simulations Workflow (PSW) Graph (WS-PGRADE)

8.3.2.1 Scientific Merit

The Cosmic Microwave Background (CMB) preserves a picture of the Universe as it was about 380 000 years after the Big Bang, and it can reveal the initial conditions for the evolution of the Universe. Planck's main objective was to measure the fluctuations of the CMB with an accuracy set by the fundamental astrophysical limits. The spacecraft charted the most accurate maps yet of the CMB.

The workflows are used to develop a web application of the Planck simulation software based on the SSP. This web application allows Astronomers to execute a large number of simulations needed to study the cosmological parameters; it helps in the challenging task of identifying and correcting instrumental and observational systematics.

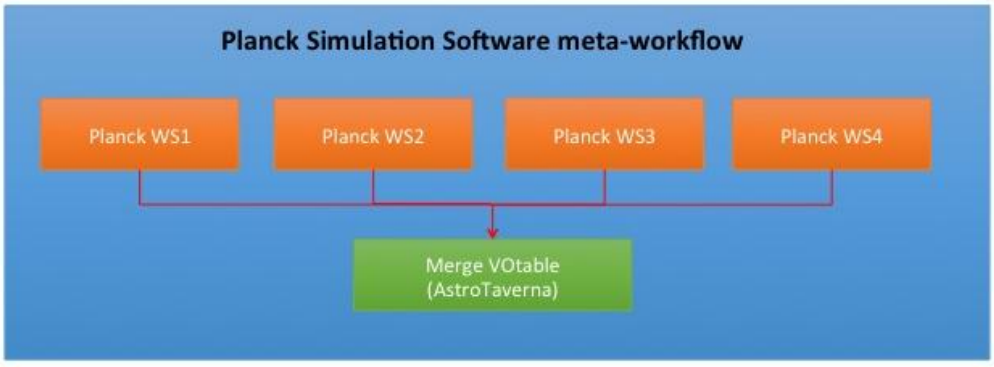


Figure 7, Planck Simulation meta-workflow

8.3.2.2 Steps

The Planck workflow consists of a very simple pipeline of different software modules. The basic steps of the pipeline are described below:

- the CMB power spectrum is created with cmbfast, starting from cosmological parameters;

- the CMB maps are built starting from the CMB power spectrum, with synfast code being part of the HEALPix package;
- the CMB is combined with foregrounds with their own frequency dependent intensities and the final sky is convolved with the beam pattern for each of the detectors considered in the simulation;
- the map is contaminated by introducing instrumental noise which is computed and added to the "observed" sky signal, therefore the TOD (Time Ordered Data) is built.

The knowledge level increases over the time, hence new details are introduced and the whole computational chain is iterated many times, even during the operative phase of the mission.

In order to speed up calculations, we can assume a perfect overlap between samples in two consecutive scan circles of the spacecraft when it remains in the same pointing position. In this way the sky signal is always the same for all the 60 scan circles corresponding to the same pointing position, so we can simulate it only once. We refer to this "fast" simulation procedure as "short" run; "long" runs instead correspond to complete simulation procedures where each scan circle is kept distinguished from the other ones.

The workflow produces as output a VOTable. The AstroTAVERNA component concatenates the various VOTables to produce a unique VOTable.

8.3.2.3 Example

A simulation starts from a set of values associated to cosmological parameters. The simulation builds an ideal sky, contaminates it and extracts new maps; a new set of parameters is obtained starting from them. As shown above, different software components contribute to build the whole pipeline run; this typical modular structure fosters the reuse of single simulation modules to build new applications and workflows.

We produced more simulations by changing the value of the Dark Matter content, then we averaged over the produced maps. The results are compared with Planck observed data.

8.3.2.4 Related Publications

- Planck Collaboration, Ade, P. A. R., Aghanim, N., Arnaud, M., et al. Planck early results. I. the planck mission. A&A, 536:A1, 2011.
- G. Taffoni et al. Enabling grid technologies for planck space mission. Future Gener. Comput. Syst., 23(2):189–200, February 2007
- Planck Collaboration, Planck intermediate results. XVI. Profile likelihoods for cosmological parameters, 2014A&A...566A..54P

8.3.2.5 Contacts

- Giuliano Taffoni: taffoni@oats.inaf.it
- Giuliano Castelli: giuliano.castelli@oats.inaf.it

8.3.2.6 Additional Information

A simulation starts from a set of values associated to cosmological parameters. The simulation builds an ideal sky, contaminates it and extracts new maps; a new set of parameters is obtained starting from these maps.

As shown above, different software components contribute to build the whole pipeline run; this typical modular structure fosters the reuse of single simulation modules to build new applications and workflows.

8.3.3 Visualization Science Case

This *visualization-oriented* workflow makes use of VisIVO tools for the production and visualization of tomographic images aimed at inspecting the cargo containers carrying high atomic number materials. The VisIVO tools provide the execution of a comprehensive collection of modules for the processing and visualization of Astrophysical and Astroparticle datasets on DCIs.

This workflow is executed in the VisIVO Science Gateway. This Science Gateway is a web-based workflow-enabled framework where a large-scale of multidimensional datasets and applications are integrated for the visualization and data filtering on Distributed Computing Infrastructures (DCIs). Advanced users are enabled to create, change, invoke, and monitor workflows; standard users, instead, are provided with easy-to-use specific web based user interfaces hiding all technical aspects of the visualization software and of the configuration and settings of the DCIs.

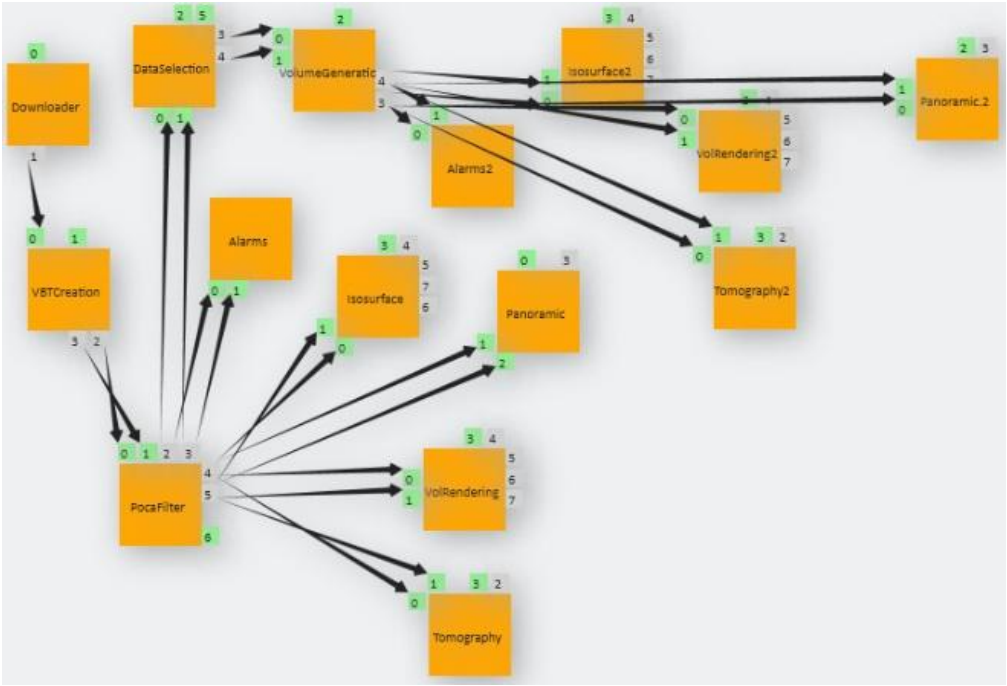


Figure 8, The VisIVO Muon Workflow Graph (WS-PGRADE)

8.3.3.1 Scientific Merit

The Muon application workflow performs a visual analysis of simulation data resulting from a scattering of cosmic radiation. The deflection of muonic particles present in the secondary cosmic radiation is the results of crossing of high atomic number materials (such as uranium or other fissile materials). This technique can provide a significant improvement compared to the traditional detection methods used so far that are based on X-ray scanners. This improvement concerns the enhanced capacity of identifying and locating illicit material, even when screens designed to mask the presence of this material are used. In this case the visualization plays a crucial role in obtaining tomographic images of a cargo container.

8.3.3.2 Steps

The datasets containing coordinates of the muon tracker planes are first uploaded to our gateway and filtered by using the Point of Closest Approach (POCA) algorithm to create a representation containing the scattering deflection of cosmic radiations. The result is then visualized using point rendering. Further processing is then applied based on user-defined

thresholds, followed by conversion into data volumes using the deflection angle field distribution by employing the 3D Cloudin-Cell (CIC) smoothing algorithm. Finally, a tomography is performed for inspection.

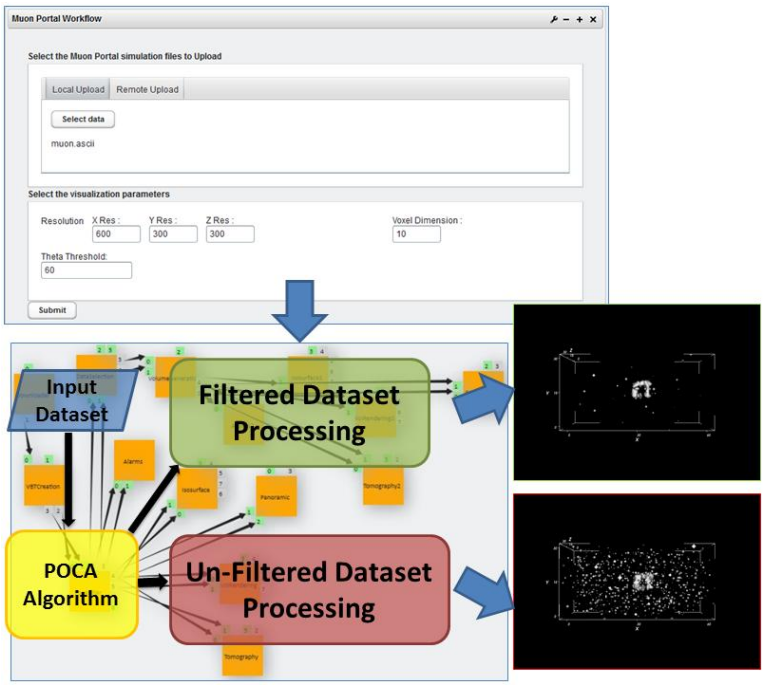


Figure 9, Muon Portal and Main Workflow Steps

8.3.3.3 Example

For this example we consider a simulation of particles in a large muon tracker consisting of 4 planes, 6 metres long and 3 metres wide, for the inspection of a container carrying high atomic number material. The illicit material is shaped with the string ”CT”.

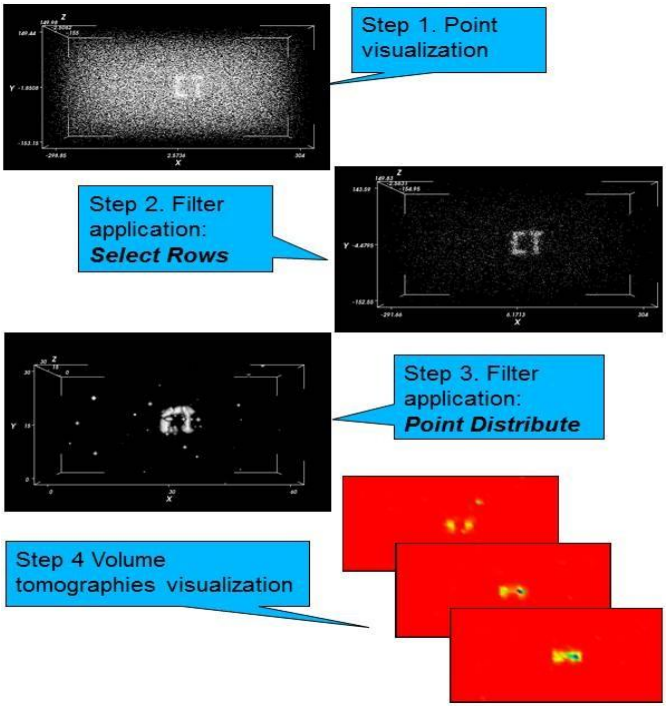


Figure 10, Muon Workflow Example Visualization Steps

The data file containing the coordinates on the muon tracker planes is first uploaded to the gateway and filtered using the POCA (Point of Closest Approach) algorithm to obtain the VBT containing the scattering deflection of cosmic radiations; the steps of the POCA algorithm are shown in Figure 11. The resulting VBT can be visualized using a point viewer as shown in the top image of Figure 10.

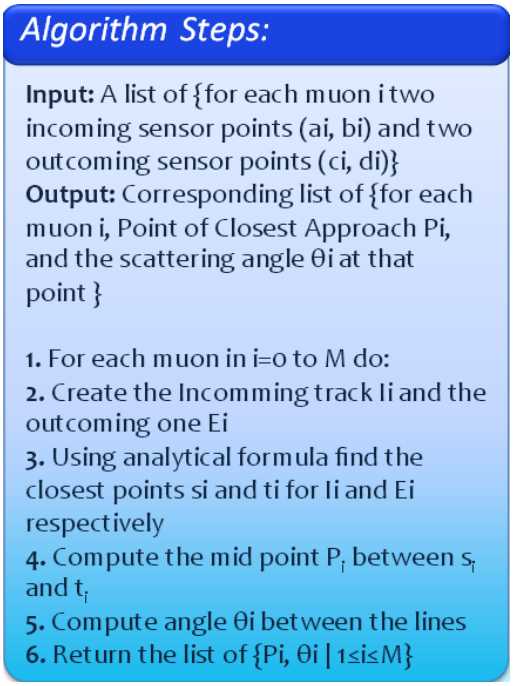


Figure 11, The steps of the POCA algorithm

The Filter portlet provides a rows filtering based on a given threshold at lower bound. The resulting VBT is then converted into a volume using the deflection angle field distribution, employing the 3D Cloud-in-Cell (CIC) smoothing algorithm, on an input defined regular mesh. The produced intermediate images are shown in steps 2 and 3 of Figure 10Not . Finally, a tomography can be performed on the produced volume VBT.

8.3.3.4 Related Publications

- 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.
- Sciacca, E.; Bandieramonte, M.; Becciani, U.; Costa, A.; Krokos, M.; Massimino, P.; Petta, C.; Pistagna, C.; Riggi, S.; Vitello, F., "VisIVO Workflow-Oriented Science Gateway for Astrophysical Visualization," Parallel, Distributed and Network-Based Processing (PDP), 2013 21st Euromicro International Conference on , vol., no., pp.164,171, Feb. 27 2013-March 1 2013
- Becciani, Ugo and Sciacca, Eva and Costa, Alessandro and Massimino, Piero and Pistagna, Costantino and Riggi, Simone and Vitello, Fabio and Petta, Catia and Bandieramonte, Marilena and Krokos, Mel, Science gateway technologies for the astrophysics community, Concurrency and Computation: Practice and Experience, DOI: 10.1002/cpe.3255

8.3.3.5 Contacts:

- Ugo Becciani: ugo.becciani@oact.inaf.it
- Eva Sciacca: eva.sciacca@oact.inaf.it
- Alessandro Costa: alessandro.costa@oact.inaf.it

8.3.3.6 Additional Information

The workflow can be accessed via VisIVO science gateway[14]. The user submits the workflow by configuring the input data files and parameters through an easy-to-use portlet interface. The workflow has a modular architecture and its building blocks can be easily reused to build other workflows.

8.4 Applications and Workflows

As mentioned in sections 8.2 and 8.3, the workflows developed during the Y2 of the project can be categorized as

- data-oriented workflows (Table 6) based on AstroTAVERNA and IVOA standards;
- lightweight building blocks (Table 7), a "library" of IVO workflows to use as re-usable modules.

Moreover, some of the Y1 workflows were revisited to add some IVO capabilities. Further details about the workflows are available in Annex A.

Name	Description	Engine	Type	Middleware	ID
BrownDwarfs Discovery	Discovery of Brown Dwarfs mining the 2MASS and SDSS databases	TAVERNA	workflow	Server/web services	5725
FindEvents InXRayAndRadio	Find events in x-ray and radio	TAVERNA	workflow	Web service	5954
HSTConeSearch	Retrieving information from HST ConeSearch and Image VO Services	TAVERNA	workflow	Web service	5958
GalaxiesSample SelectionResearch Object	Galaxies Sample Selection Research Object	TAVERNA	workflow	Web service	5960
MuonWf	VISIVO MUON	WS-PGRADE	workflow	gLite/server	5103

Table 6, Data oriented workflows implemented in Y2 by the Astrophysics Community

8.5 Applications Usage

All the ported workflows of the second year are used in the same way. There are two possible access modes that are being developed and used in the project:

- Through the **SHIWA Simulation Platform (SSP)**. Developers choose this access mode during the porting and testing phase. The SSP has been connected to gLite middleware in order to execute jobs on Grid.
- Through **science gateways** by means of a customized **user interface**. This access mode is preferred by the scientists to run the applications. The A&A science gateways can be used by advanced users who are able to design new workflows and/or combine together existing workflows or meta-workflows modules. Standard users connect to web applications developed as Java portlets that hide the complexity of the workflows and meta-workflows. Standard users interact with the web interface to configure and execute the application and monitor the execution status.

Workflow developers benefit from the IVO meta-workflows library to develop complex workflows that include data access and computations (typically data reduction and analysis

[14] http://visivo.oact.inaf.it:8080

operations). WS-PGRADE meta-workflows can be executed from the same interface as WS-PGRADE workflows from the SSP.

The A&A community mainly uses the SHIWA Simulation Platform as development and test environment, while dedicated science gateways are set up as execution environments to exploit specific Science Cases.

Name	Description	Engine	Type	Middleware	ID
VOTables Concatenation	Concatenates several VOTables into one	TAVERNA	Workflow	server	5726
FileCreation FromTemplate AndVOTable	Create configuration files from a template and a VOTable	TAVERNA	Workflow	server	5729
Astronomical NameTo Equatorial Coordinates Resolver	Astronomical object name to equatorial coordinates Resolver	TAVERNA	Workflow	server	5140
RunScripts FromAVO TableColumn	Run scripts from a column in a VOTable	TAVERNA	Workflow	server	5730
CreateVOTableF romEllipse Results	Create VOTable from ellipse results	TAVERNA	Workflow	server	5142
AddColumns ToVOTable From Sextractor Execution	Add columns to a VOTable resulting from the execution of sextractor	TAVERNA	Workflow	sever	5143
ExtractColumnF romVOTable IntoList	Extract a column from a VOTable into a List	TAVERNA	Workflow	server	5144
SplitVOTable IntoItsValues	Split VOTables into its values	TAVERNA	Workflow	server	5952
ExtractContentO fColumns FromVOTables	Extract content of columns from VOTables	TAVERNA	Workflow	server	5953
VOTableOf NEDImages FromAList OfObjects	VOTable and NED Images from a list of Objects	TAVERNA	Workflow	sever	5955
ConeSearch QueryTo VOService	Perform a ConeSearch query to a VO Service	TAVERNA	workflow	server	5959

Table 7, Lightweight building blocks implemented in year two from the Astrophysics Community

9 Computational Chemistry

The Molecular Simulation Grid (MoSGrid) is a German project that aims at easing the access and use of molecular simulations in Computational Chemistry in a grid environment. The Computational Chemistry is an established discipline in natural sciences; it targets modelling and analysing three-dimensional molecular structures. Important application domains are molecular dynamics, quantum chemistry, and docking. Each of these domains consists of a diverse set of scientific simulation programs and data flows. The data flows of the chemical simulations consist of many possible steps, including file transfers, data conversions, and molecular analyses. Hereby, the state of the art available simulation codes, hand in hand with today's high performance computing infrastructures, allow molecular simulations to solve increasingly complex scientific questions. Therefore, more and more scientists are using these tools.

However, even today's most powerful simulation instruments still have limitations, especially due to the design of the user interfaces. Many sophisticated tools are command-line driven and not supported by a graphical user interface. As a consequence, the new users have to become familiar not only with the large number of methods and chemical theories, but also with the use of the codes and the handling of the data flows. To lower the hurdle of using these programs, intuitive and data driven user interfaces are paramount.

MoSGrid offers a gUSE/WS-PGRADE based science gateway that allows an easy access to complex molecular simulations. The included web-based graphical user interface allows a simulation code independent setup of simulation workflows that are submitted through the UNICORE grid middleware to the underlying clusters. Every user can apply commonly used metadata enriched workflows that are available in recipe repositories. The metadata description allows an efficient search for the required workflows by a description of the underlying dataflow. This lowers the hurdle for applying computational chemistry methods even for novice users.

9.6 Technical Background

The **MoSGrid** science gateway has been developed on top of **WS-PGRADE** (Web Services Parallel Grid Runtime and Developer Environment), which employs the portal framework Liferay and forms the highly flexible user interface of **gUSE** (grid User Support Environment). The MoSGrid portal offers a graphical workflow system.. Commonly used simple and complex workflows can be stored in recipe repositories and are made available for every user. As underlying middleware, **UNICORE** has been chosen after a requirement analysis. It offers a complete stack of tools; a graphical user interface allows to create jobs and workflows and to submit them to a UNICORE grid that can consist of several clusters. UNICORE middleware services manage jobs and authenticate and authorize users. A service running on logins nodes of clusters communicates with these to run jobs for users.

In the MoSGrid project a new submitter for UNICORE was developed and contributed to gUSE. It allows the submission of workflow tasks to UNICORE grids. In this way the jobs can be easily distributed to clusters all over Europe. The submitter also includes functionality to index metadata. For this the UNICORE metadata service is instructed to automatically index available metadata at the end of a workflow. This makes the metadata searchable for later use (see more details in the deliverable D5.3 – Requirements for domain semantic data and workflow description).

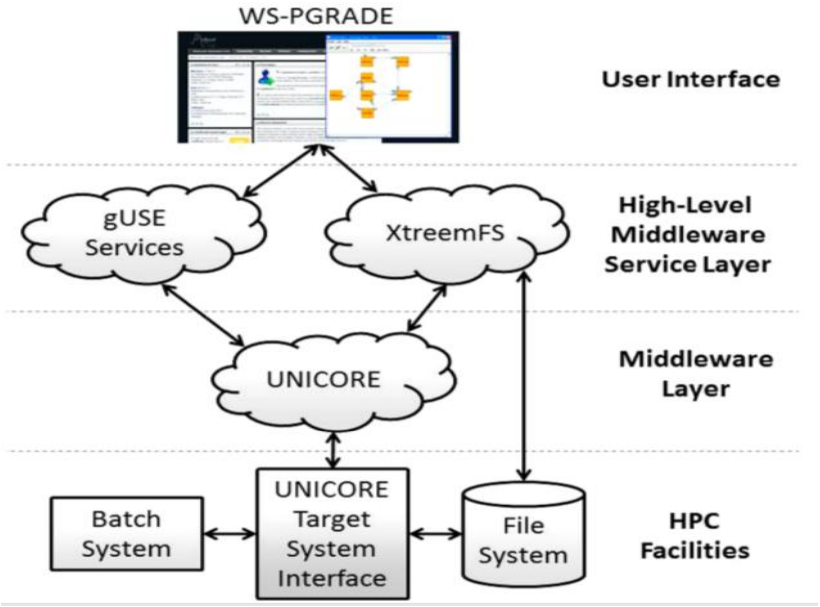


Figure 12, Technical Background of the Computational Chemistry Community

Furthermore WS-PGRADE, as the graphical user interface to gUSE, was extended to support the UNICORE incarnation database (IDB). On the one hand users are enabled to easily select tools installed on clusters to be used in workflows. Only tools available on at least one cluster can be selected. On the other hand jobs will only be submitted to a cluster where the chosen tools are available. The application does not have to be transmitted to the cluster; instead, already installed applications are used. The user also does not have to know where the applications are installed on a cluster or on which cluster it is installed. The WS-PGRADE graphical user interface is indeed utilizable by chemists who have no informatics expertise. A graph editor offers a visual access to the design of the workflow parts (tasks, input and outputs ports, connections). The subsequent “real” workflow definition is designed in an almost intuitive way of clicking through the steps.

The MoSGrid science gateway enables the user to easily find data again. This functionality consists of a search field where terms are entered. When a term matches metadata associated to data, this data is displayed and can be selected for further analysis.

9.7 Workflow Usage in the Community

For the second year, we have again chosen applications which are representative for every domain of MoSGrid (Molecular Dynamics, Docking, Quantum Chemistry). As generic approach, we have decided to use open-source simulation software.

Basically, we built up workflows within the MoSGrid environment and then port them to the SHIWA Repository for storage and sharing. In parallel, we use Galaxy workflows in the Galaxy portal (http://flavus.informatik.uni-tuebingen.de:9090/) and port them to the SHIWA Repository as well. Since UNICORE as workflow engine and middleware is highly practical, it has found wide use in the Computational Chemistry community. The UNICORE workflows cannot be ported to SHIWA at the moment, but as they represent a different approach to similar quantum chemical problems they shall be summarized here as well.

Since Chemists needed some time to become familiar with the concept of meta-workflows, we herein summarize new workflows as well as meta-workflows and even meta-meta-workflows.

Note that in MoSGrid we chose to document the workflows, both in this deliverable as in the SHIWA Repository, using abstract visual representations, and sometimes screenshots of the

workflows implemented in WS-PGRADE. This choice was motivated by the goal of providing meaningful documentation for the users of this community, which are scientists interested in the functionality implemented by the workflow steps rather than implementation details, as the screenshots reveal.

9.8 Science Cases

This community developed three science cases.

Name	Description	Workflow IDs
Spectroscopic Analysis	Explore the spectroscopic characteristics of a molecule	5739
Spectroscopic Benchmarking	Calculation of optimized geometries, molecular orbitals, population analyses, frequencies, or optical absorptions.	5745
Population UNI	Apply various population schemes to better electronic understanding of the molecules.	http://www.erflow.eu/documents/388575/775509/Application.PopulationUNI.pdf

Table 8, Computational Chemistry Science Cases

9.8.1 Spectroscopic Analysis Science Case

A highly useful quantum chemical Science case is the so-called spectroscopic analysis. After a first geometry optimization of the desired molecule, several further simulations are performed which serve for a spectroscopic analysis. Chemists describe this in a rather complex workflow, which comprises a multitude of consecutive and subsequent steps. First of all, the geometry of the desired molecules needs to be energetically optimized. In real-life most inorganic chemists look at molecules which possess at least 50-200 atoms. A quantum chemist wants to explore the spectroscopic properties of such molecules. This can be on the one hand the vibrations of the molecule (IR and Raman frequencies) and on the other hand UV/Vis spectra of such a molecule. The vibrations require a so-called frequency calculation. In case of only positive vibrations, the molecule geometry represents a true minimum. UV/Vis spectra with good accordance to experimental data are obtained by time-dependent density functional theory calculations (TD-DFT).

9.8.1.1 Scientific Merit

The spectroscopic analysis of selected molecules allows a better interpretation of experimental spectroscopic data and helps to an identification of highly reactive chemical species. The species can then be further developed towards sustainable catalysis.

9.8.1.2 Steps

When dissecting the single steps of the idea described above, we identified that there are smaller workflows of fundamental quality (such as the optimization workflow) which are embedded in the larger entity. The dissection followed the principle of identifying small tasks which can be reused within other workflows. Hence, one can define several small workflows as part of a larger meta-workflow as depicted in Figure 13.

The workflow dissection provides with the insight that the first workflow is a simple geometry optimization (opt WF). Such a basic workflow can be reused in many more applications. The subsequent workflows are similar to each other: a converter script extracts the output geometry from the optimization output and combines it with blank input files (i.e. just lacking input coordinates) with the corresponding keywords for frequency calculations (leading to a Freq WF), time-dependent DFT (TD-DFT WF), population analyses (pop WF) and

subsequent calculations in solvents (Solv WF). All these small workflows in Figure 13 are highly valuable since they can be reused in larger quantum chemical workflows. The whole systems gains flexibility as the small workflows can be freely combined to new meta-workflows.

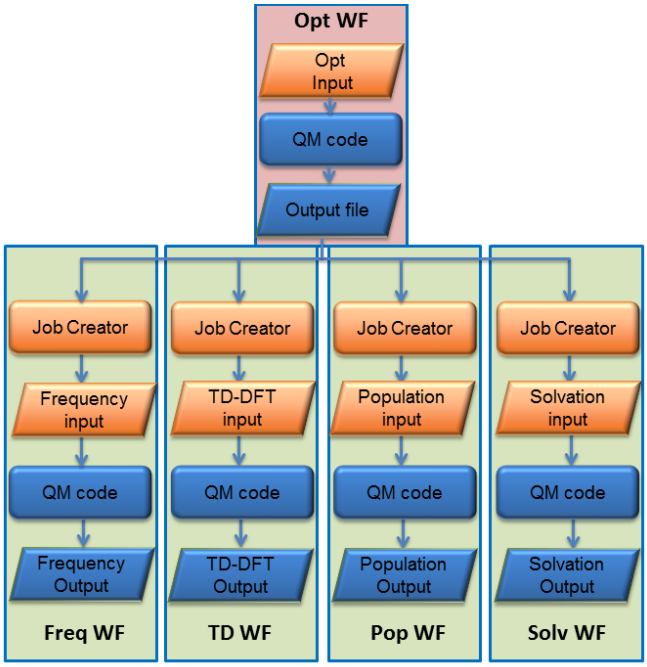


Figure 13, Spectroscopic workflow after dissection into basic workflows

9.8.1.3 Example

Understanding of the formation of peroxo (Figure 14) and the isomeric oxo cores as well as their distinct reactivity relies on comprehensive orbital analyses. First the geometry is optimised then TD-DFT and all other steps are performed. Figure 15 shows as example the optical spectra.

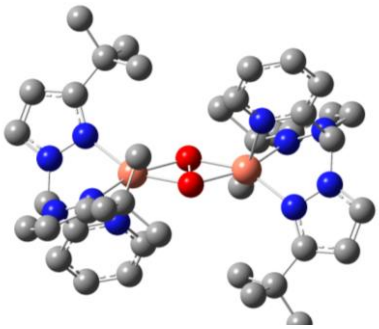


Figure 14, Peroxo dicopper bis(pyrazolyl)methane complex (grey: carbon atoms, blue: nitrogen atoms, red: oxygen atoms, copper: copper atoms

Time dependent-DFT calculated spectra predict the optical spectrum with the four LMCT bands at 340 nm, 366 nm, 381 nm and 547 nm (Figure 15) in good accordance to the experimental spectrum.

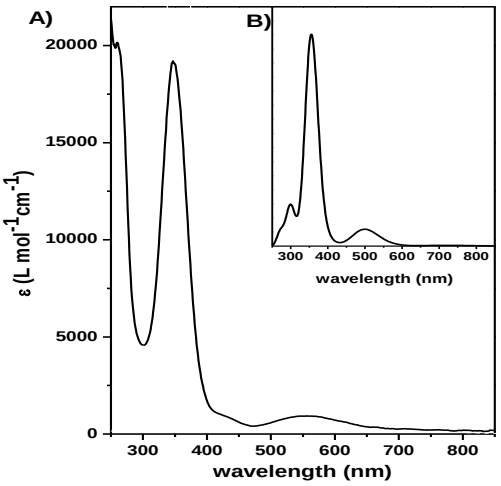


Figure 15, UV/Vis spectra of the peroxo dicopper bis(pyrazolyl)-methane complex (A: experimental, B: calculated with Gaussian09, TD/DFT, 80 states, TPSSh, 6-31g(d))

9.8.1.4 Related Publications

1. S. Herres-Pawlis, A. Hoffmann, S. Gesing, J. Krüger, A. Balasko, P. Kacsuk, R. Grunzke, G. Birkenheuer, L. Packschies, User-Friendly Workflows in Quantum Chemistry, *CEUR Workshop Proceedings* **2013**, *993*, Paper 14.
2. S. Herres-Pawlis, A. Hoffmann, A. Balasko, P. Kacsuk, G. Birkenheuer, A. Brinkmann, L. de la Garza, J. Krüger, S. Gesing , R. Grunzke, G. Terstyansky, N. Weingarten, Quantum chemical metaworkflows in MoSGrid, *Concurrency Computat.: Pract. Exper.* **2014,** in print.
3. A. Hoffmann, R. Grunzke, S. Herres-Pawlis, Insights into the influence of dispersion correction in the theoretical treatment of guanidine-quinoline copper(I) complexes*, J. Comp. Chem*. **2014**, *35*, 1943–1950.

9.8.1.5 Contacts

- Dr. Alexander Hoffmann, e-mail: alexander.hoffmann@cup.uni-muenchen.de
- Prof. Dr. Sonja Herres-Pawlis, e-mail: sonja.herres-pawlis@cup.uni-muenchen.de

9.8.2 Spectroscopic Benchmarking Science Case

The full simulation of molecular structures including the electronic structures comprises the calculation of optimized geometries, molecular orbitals, population analyses, frequencies, or optical absorptions. The combination of all these tasks as basic workflows into a meta-workflow improves the simulation enormously (see “Spectroscopic analysis”). With regard to real-life systems, the simulation has to tackle issues such as antiferromagnetic coupling between copper atoms, correct description of the coordination sphere and multiple conformations of the whole molecule. Methodologically, density functional theory is most appropriate here due to size of the system and investigated questions. Hence, the spectroscopic workflow needs to be performed several times for an array of functionals and basis sets which have to be tested for the ultimate structural and optical description with regard to experimental data.

Furthermore, the spectroscopic meta-workflow can be combined into a new type of workflow called meta-meta-workflow with all being implemented in WS-PGRADE. Figure 16 shows four spectroscopic meta-workflows are combined into a meta-meta-workflow after performing a basic optimization. This basic optimization (basic opt WF) serves as pre-optimization step which saves calculation time in all subsequent optimizations included in the spectroscopic

workflows (specX WFs). A meta-meta-workflow saves a lot of time in this application - more than a normal meta-workflow.

9.8.2.1 Scientific Merit

The spectroscopic benchmarking needs to be done for every molecule of a new class of molecules. Afterwards, the experience made in these analyses can be transferred to further members of the regarded class. This saves a lot of job definition time for the researcher.

9.8.2.2 Steps

The first input file is a opt.nw file for a basic optimisation simulation. The output of the first basic WF is a opt.out file which is parsed for the geometry by the subsequent workflows. They combine this geometry data with prepared nw-input-files for the subsequent freq, TD, Mull and solv jobs. As final output, multiple sets of output files freq.out, TD.out, Mull.out and solv.out are obtained (Figure 16).

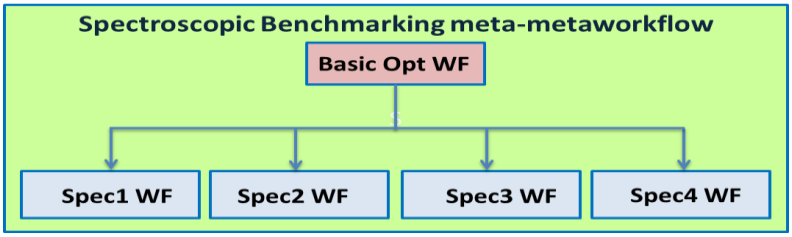


Figure 16. NWChem meta-meta-workflow for spectroscopic benchmarking

9.8.2.3 Example

See section 9.3.1.4 for an example.

9.8.2.4 Related Publications

1. Herres-Pawlis, S., Hoffmann, A., de la Garza, L., Krüger, J., Grunzke, R., Gesing, S., Weingarten, N., and Terstyansky, G., Meta-metaworkflows for Combining Quantum Chemistry and Molecular Dynamics in the MoSGrid Science Gateway, IWSG **2014** (6th International Workshop on Science Gateways), June 2014, Dublin, Ireland. Accepted.
2. S. Herres-Pawlis, A. Hoffmann, A. Balasko, P. Kacsuk, G. Birkenheuer, A. Brinkmann, L. de la Garza, J. Krüger, S. Gesing , R. Grunzke, G. Terstyansky, N. Weingarten, Quantum chemical metaworkflows in MoSGrid, *Concurrency Computat.: Pract. Exper.* **2014,** in print.
3. A. Hoffmann, R. Grunzke, S. Herres-Pawlis, Insights into the influence of dispersion correction in the theoretical treatment of guanidine-quinoline copper(I) complexes*, J. Comp. Chem*. **2014**, *35*, 1943–1950.

9.8.2.5 Contacts

- Dr. Alexander Hoffmann, e-mail: alexander.hoffmann@cup.uni-muenchen.de
- Prof. Dr. Sonja Herres-Pawlis, e-mail: sonja.herres-pawlis@cup.uni-muenchen.de

9.8.3 Population UNI Science Case

Besides the standard natural bond orbital analysis implemented in many quantum chemical codes, the state-of-the-art is represented by the NBO6.0 analysis. This is a standalone programme which needs special input. Hence, after the basic optimisation with NWChem a special input file for the subsequent single point job has to be generated by a script. This input file then goes into the single point calculation using NWChem (sp files). The output files of these calculations are then the input files for NBO6, AOMix and AIM. AOMix and AIM are small but commercial programmes where a working group licence is needed. They provide

with different kinds of population analyses which allow for different orbital dissections, electronic analyses, and calculation of bonding parameters and so on.

9.8.3.1 Scientific Merit

Orbital analyses depend on the chosen model for the question to the molecule. Hence, various population schemes should be applied to avoid misinterpretations. This study leads to a better electronic understanding of the regarded molecules.

9.8.3.2 Steps

The first input file is an opt.nw file for a basic optimisation simulation. The output of the first basic WF is an opt.out file which is parsed for the geometry by the subsequent script which generates the input files for NWChem jobs which generate the input files for the population analyses of the other codes (Figure 17).

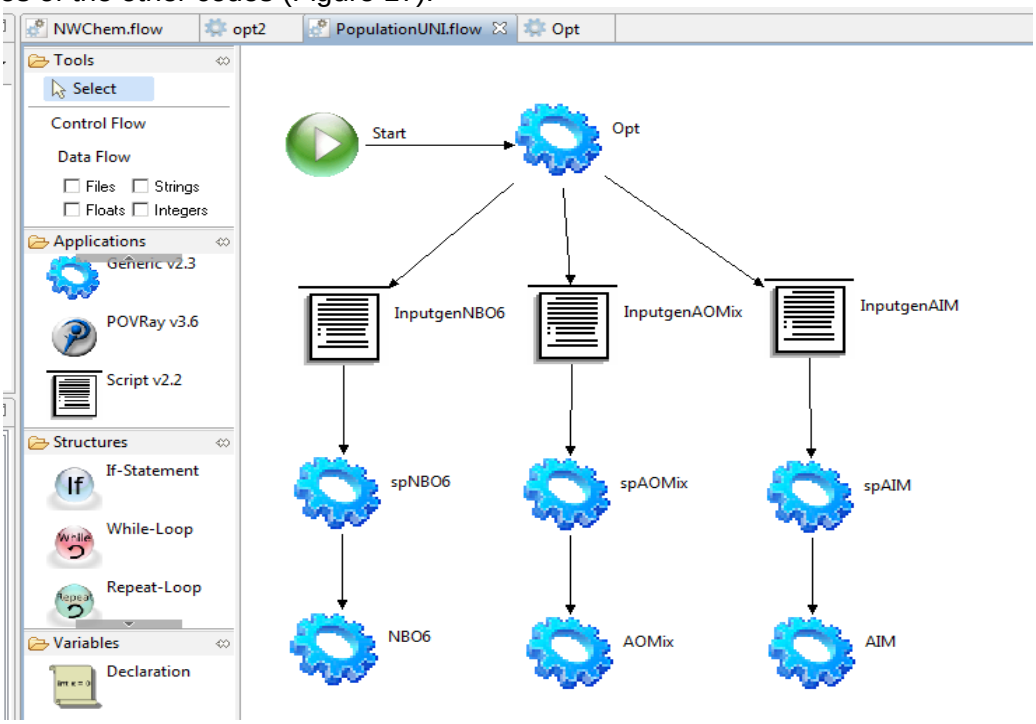


Figure 17. NWChem workflow for full population analysis using NWChem and NBO, AOMix and AIM in UNICORE

9.8.3.3 Example

For a series of bis(pyrazolyl)methane transition metal complexes we performed NBO analyses. The NBO analysis yields also the charge transfer energies (by means of second order perturbation theory) for the donation from the pyrazolyl/pyridinyl units to the metal ions.

The NBO calculation shows that the N_{py} donor atoms possess a more negative charge than the N_{pz} donor atoms independent of coordinating a metal or not. Furthermore, the relative basicities of the donor functions of the ligand $HC(Pz)_2(Py)$ have been determined by DFT. The protonation of the pyridinyl donor is 9 kcal/mol more favourable than the protonation of the pyrazolyl donor. This result agrees to the known pK_B values of pyrazole (11.5) and pyridine (8.8). So the pyridinyl donor is the stronger base. Basicity and donor strength often correlate strongly, so in complexes, the pyridinyl function might also act as stronger donor. In this complicated situation, the donor competition between pyrazolyl and pyridinyl donors is very close and can be influenced by subtle varieties such as chelate bite, spin state and Jahn-Teller distortion.

By means of a NBO analysis for all complexes and comparative complexes, partial charges, charge transfer energies and hybridisation of the donating atoms have been calculated. Hereby, the donor competition has been elucidated and set in comparison to experimental structural data. In general, with only small special exceptions, we find that the pyrazolyl donors donate more strongly to the metals with shorter bonds although the pyridinyl donors are more basic. This delicate bias towards pyrazolyl as stronger donor can easily be disturbed by the change of coordination geometry, spin state and Jahn-Teller effects.

9.8.3.4 Related Publications

1. A. Hoffmann, U. Flörke, S. Herres-Pawlis, Insights into Different Donor Abilities Within Bis(pyrazolyl)pyridinylmethane Transition Metal Complexes, Eur. J. Inorg. Chem. 2014, 2296-2306.
2. A. Hoffmann, R. Grunzke, S. Herres-Pawlis, Insights into the influence of dispersion correction in the theoretical treatment of guanidine-quinoline copper(I) complexes*, J. Comp. Chem*. **2014**, *35*, 1943–1950.

9.8.3.5 Contacts

- Dr. Alexander Hoffmann, e-mail: alexander.hoffmann@cup.uni-muenchen.de
- Prof. Dr. Sonja Herres-Pawlis, e-mail: sonja.herres-pawlis@cup.uni-muenchen.de
- Dr. Jens Krüger, e-mail: Krueger@informatik.uni-tuebingen.de

9.9 Applications & Workflows

Similarly to Year 1 the chosen applications represent every domain of MoSGrid (Molecular Dynamics, Docking, Quantum Chemistry). After developing simple (or basic or building block) workflows in Year 1, in Year 2 first, we focused on meta-workflows combining basic workflows (or workflow building blocks). Next, we developed a few meta-meta-workflows. The meta-workflows and meta-meta-workflows have proven their high re-usability in many contexts.

NWChem has proven to be a useful highly scalable code for Quantum Chemistry as well as Molecular Dynamics applications. Hence, we have created further NWChem basic workflows as building blocks. These basic workflows can be transferred to Gaussian or ORCA as well.

The following table summarizes the meta-workflows and meta-meta-workflows with their sub-workflows. Further details about the workflows are available in Annex B.

Remark: When not differently stated, WS-PGRADE was used for workflow implementation.

Meta-workflow	Description	Sub-workflow	Workflow ID
Spectroscopic analysis	Explore the spectroscopic characteristics of a molecule	NWChem_basic NWChem_freq NWChem_TD NWChem_Mull NWChem_solv	3958 4206 4751 4753 4752
Spectroscopic benchmarking	Explore the spectroscopic characteristics of a molecule with more functionals/basis sets	NWChem_basic NWChem_Spectros-copic analysis	3958 5739
Transition state analysis	Find a reaction transition state and analyse its frequencies together with the spectroscopic properties	NWChem_TransitionStateSearch NWChem_freq NWChem_Spectros-copic analysis	4913 4206 5739

QM-MD	Optimise the protein scaffold by molecular dynamics and optimise then the metallo-active center by quantum mechanics together with a spectroscopic analysis	Gromacs_protein-equilibration NWChem_Spectros-copic analysis	4354 5739

Table 9, Overview of metaworkflows and their subworkflows (all workflows were built up using WS-PGRADE)

In the second year, the workflows summarized in Table 10 were built up and ported.

Name	Description	Engine	Middle-ware	Type	ID
MD solvation	Solvate a molecule within a molecular dynamics scheme	WS-PGRADE	UNICORE	workflow	5748
Galaxy MD	Solvate a molecule within a molecular dynamics scheme in Galaxy	Galaxy	UNICORE	workflow	5755
AutoDockVina ASDD	Dock a molecule into a receptor with AutoDockVina	WS-PGRADE	UNICORE	workflow	5747
Galaxy Docking	Dock a molecule into a receptor in Galaxy	Galaxy	UNICORE	workflow	5743
Spectroscopic analysis	Explore the spectroscopic characteristics of a molecule	WS-PGRADE	UNICORE	meta-workflow	5739
Spectroscopic benchmarking	Explore the spectroscopic characteristics of a molecule with more functionals/basis sets	WS-PGRADE	UNICORE	meta-meta-workflow	5745
Parameter sweep	Benchmark a molecule using larger arrays of functionals and basis sets	WS-PGRADE	UNICORE	meta-workflow	--
Transition state analysis	Find a reaction transition state and analyse its frequencies together with the spectroscopic properties	WS-PGRADE	UNICORE	meta-workflow	5751
TD-UNI	Calculate the optical response of a molecule	UNICORE	UNICORE	workflow	
SpecWSUNI	Explore the spectroscopic characteristics of a molecule	UNICORE	UNICORE	workflow	
PopulationUNI	Apply various population schemes to better electronic understanding of the molecules.	UNICORE	UNICORE	workflow	
Galaxy QM	Optimise a molecule quantum chemically in Galaxy	Galaxy	UNICORE	workflow	5754
QM-MD	Optimise the protein scaffold by molecular dynamics and optimise then the metallo-active center by quantum mechanics together with a spectroscopic analysis	WS-PGRADE	UNICORE	meta-meta-workflow	5752

Table 10, Overview of workflows in Year 2

9.10 Applications Usage

The ported workflows are used by the different types of users in different ways. There are five possible access modes that are being developed in the project:

- Through the **SHIWA Simulation Platform (SSP)**. This mode is preferred by developers during the porting and testing phase. During the second year, the UNICORE IDB was successfully connected to the SSP such that WS-PGRADE jobs are able to run on UNICORE resources.
- Through the **MoSGrid Developer Interface**. This mode is preferred by workflow developers who wish to combine workflows and develop new ones. The MoSGrid portal enables to build up workflows using WS-PGRADE in a white-box approach. The white-box approach proved to be superior to the black-box approach for building workflows since the black-box workflows were too error-prone. From the MoSGrid developer interface the workflows can be easily ported to the SHIWA repository via one click at the general workflow list of the MoSGrid developer. Only the SHIWA repo password and the domain information have to be added. Afterwards more information can be added to the ported workflow in the SHIWA repo. This works really straightforward.
- Through the **MoSGrid User Interface** divided after domains. This mode is preferred by end-users who just wish to have access to the domain-specific portlets and the workflows accessible to the standard users. They do not see the technical details but a user-friendly description of the workflow and its performance.
- Through the **UNICORE User Interface**. This mode is preferred by users who use simple workflows and fast hardware behind the scenes.
- Through the **Galaxy portal**. This mode is preferred by docking users who are used to this workflow engine and their portal

10 Heliophysics

Heliophysics is the branch of physics that investigates the relationship among the various bodies of the Solar System. More precisely: it investigates how the Sun influences the Heliosphere. The investigative domain of Heliophysics is thus the entire Solar System, an extremely vast “experimental space” where phenomena propagate from the Sun to the outer bodies. Data is gathered from a multiple of sensors on board satellites and on the surface of the planets. Successful investigation in Heliophysics entails coordinated analysis of data gathered across the entire Solar System regarding phenomena that evolve in space and time, a complicated problem that is tackled with mathematical abstractions called propagation models. The Heliophysics community has started using workflows with the HELIO FP7 project in 2009 and has since then adopted a cross platform approach using TAVERNA and WS-PGRADE workflows.

Workflows are particularly useful to the Heliophysics community as they address one of its salient characteristics, that of a short life-span of its science cases. When a workflow is developed to investigate a specific phenomenon, it may lack usability for other events and thus the need to develop highly re-usable workflows in the same fashion of object-oriented development is paramount to success. To assess the usability and usefulness of the developed workflows, we have built three meta-workflows (two with TAVERNA and one with WS-PGRADE) that address Science Cases.

The workflows ported during the second year of the ER-FLOW project follow this multi-layered approach to foster re-usability of single components. The services orchestrated by the workflows are those developed and maintained by the HELIO (http://www.helio-vo.eu/).

10.1 Technical Background

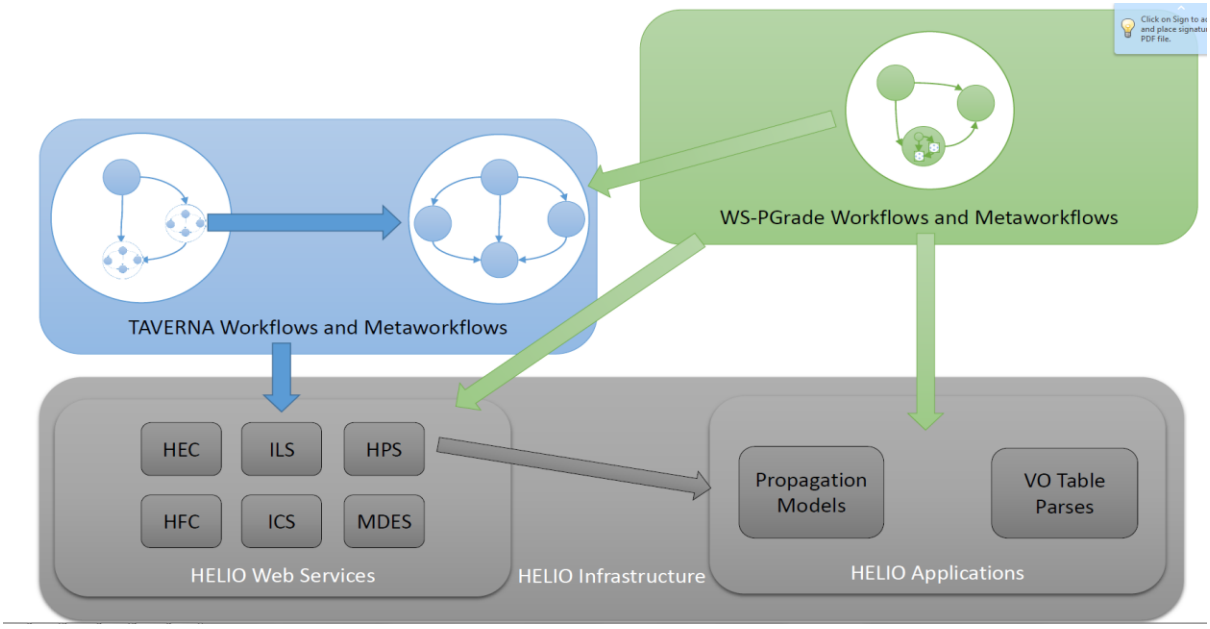


Figure 18, Execution environment of the Heliophysics workflows

The service architecture where the workflows are executed is the result of three different EU-funded projects:

- HELIO (http://www.helio-vo.eu/) provides and maintains the web-services that are orchestrated by the workflows,
- SCI-BUS (http://www.sci-bus.eu/) provides and maintains the HELIOGate science gateway where the workflows are designed and executed, and,

- ER-flow, for the SHIWA (https://www.shiwa-workflow.eu/) that provides the interoperability platform that allows the execution of workflows across different technologies (TAVERNA and WS-PGRADE for the Heliophysics community). ER-flow also maintains the generic SHIWA Simulation Platform that is also used to develop and execute Heliophysics workflows.

The workflows are arranged in the architecture illustrated in Figure 18 and they are executed in the infrastructure illustrated in Figure 19..

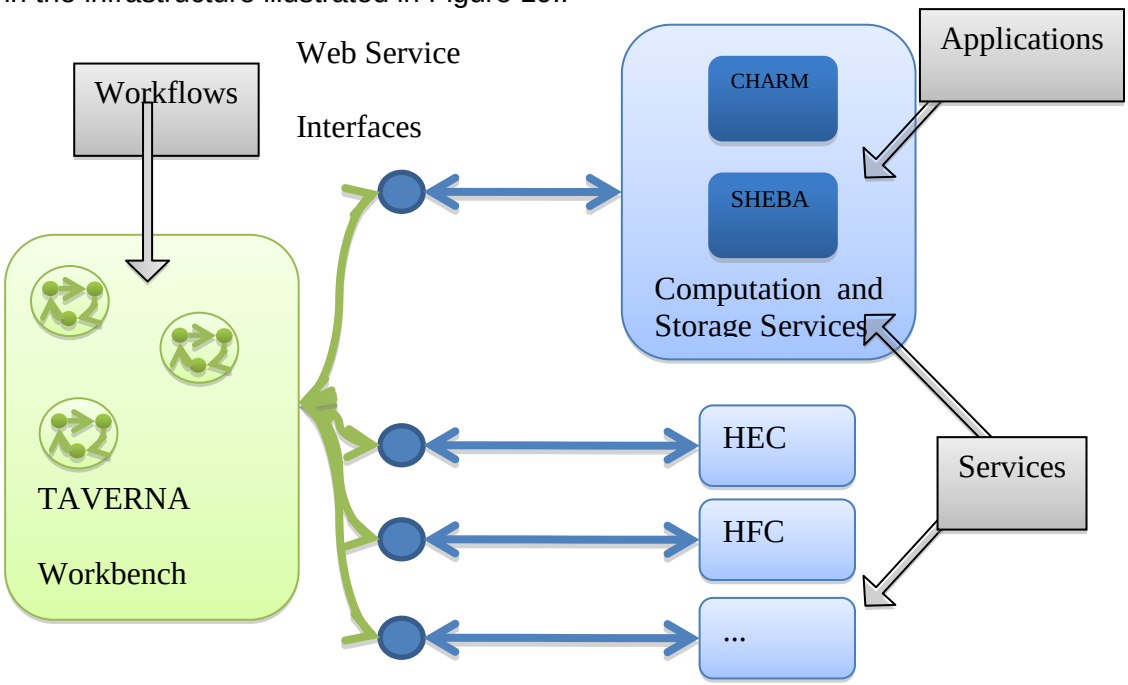


Figure 19, Services, Workflows and Application of HELIO

The workflows rely on service oriented architecture. These resources can be roughly divided in applications, services and workflows, as described in the figure above:

- **Applications**: programs or routines that can be used in standalone modality or as part of a wider workflow orchestration. Among the various applications used in Heliophysics, two kinds are most common and useful: Feature extraction applications and propagation models.
  - *Feature extraction* codes are used to find relevant features from raw data. Among these images processing codes are very common; feature extraction applications usually provide input to catalogues and lists of metadata. **SMART** is a feature extraction code that spots active regions from magnetograms of the surface of the sun. It is a set of routines written in the IDL language and it needs the SSW libraries written in IDL.
  - *Propagation models* describe the movement of physical features throughout the solar system; they are used to find cause-effect relations between events spotted at different places and different times in the Solar System. **SHEBA** is a feature extraction code that spots active regions from magnetograms of the surface of the sun. It is a set of routines written in the IDL language and it needs the SSW libraries written in IDL.
- **Services**: programs or routines that are accessible through web services interfaces.
  - **The Heliophysics Event Catalogue (HEC)** is used to query about 60 different catalogues of events.
  - **The Heliophysics Feature Catalogue (HFC)** is used to query catalogues of solar features including filaments, coronal holes, sunspots and active regions.

- **The Context Service (CXS)** is used to determine the context in the heliosphere in a given time frame. It can trace light curves, plot flare locations and plot the Parker Spiral.
- **The Instrument Capability Service (ICS)** is used to determine the capabilities of an instrument such as the observable entities and the observing domains.
- **The Instrument Location Service (ILS)** is used to determine the position in the solar system of an instrument.
- **The Data Provider Access Service (DPAS)** is used to access diverse data sources regardless of the access protocol (http, ftp, etc.).
- **The Data Evaluation Service (DPAS)** is used to evaluate and display numerical (series) data.

- **Workflows**: orchestration patterns of Applications and Services. All the workflows ported during this second year have been developed with **TAVERNA**.
- **Middleware and Resources**: All the ported workflows do not access directly middleware or computational and storage resources. The applications CHARM and SHEBA are executed on specific HELIO services, the HELIO Processing and Storage Services (**HPS** and **HSS**), which expose a web service interface to computation and storage resources
- **Certificates and Virtual Organizations:** As all ported workflows do not access any Grid Resource, therefore so far there is no need for grid certificates or affiliation with any Virtual Organization.

10.2 Workflow Usage in the Community

After Year 1, we have acknowledged that the methodology followed so far lacked in flexibility and re-usability and we have changed it accordingly. The workflows are now designed around a multi-layered concept based on meta-workflows and workflow interoperability as illustrated in Figure 20.

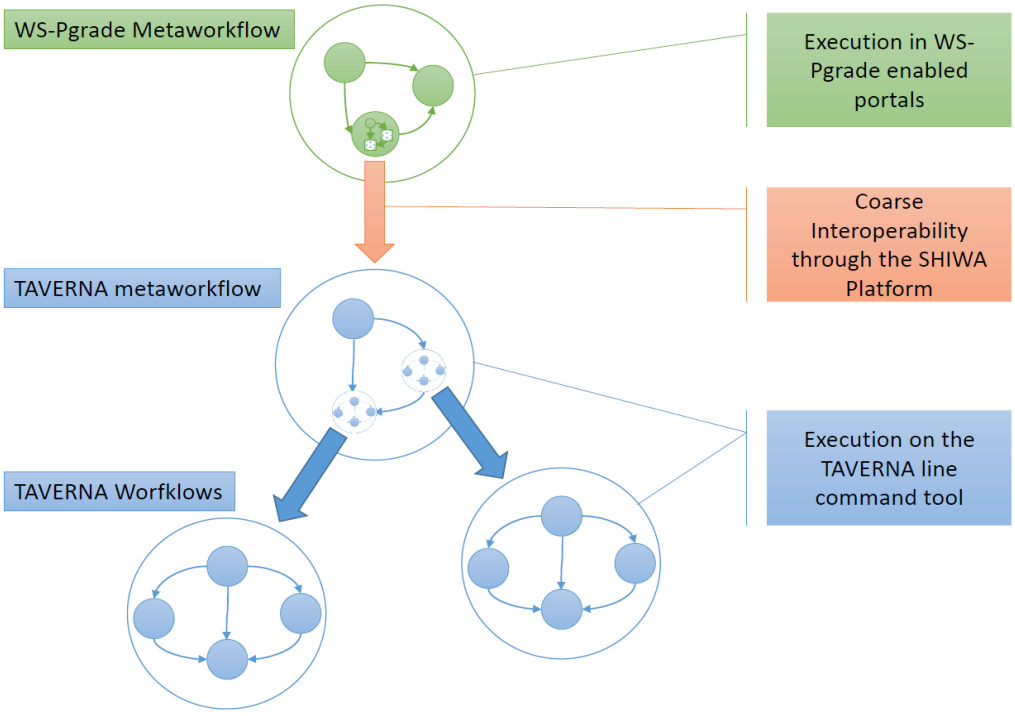


Figure 20, Metaworkflows and workflows interoperability in Heliophysics

Our aim is to develop a suite of workflows and meta-worfklows that must be as easy as possible in their use, that can be re-used to minimize user's effort and that can be used successfully to tackle real science cases. In order to achieve our aims we have divided the workflows as belonging to two main categories: science meta-workflows and basic workflows.

- **Basic Workflows** that usually define a simple task (defined as a Use Case) that can be represented by a simple application (such as a script or other executables executed locally on the TAVERNA engine or on a Distributed Computation Infrastructure for WS-PGRADE workflows) or by the remote invocation of a Web Service. The HELIO services expose two interfaces: we use the SOAP interface for the TAVERNA workflows and the REST interface for the WS-PGRADE workflows.
- **Science Workflows** that implement Science Cases (the definition of a scientific challenge) by composing different Basic Workflows. When they are developed in TAVERNA, use TAVERNA Basic Workflows, but when they are developed in WS-PGRADE, they can either use WS-PGRADE or invoke TAVERNA through the SHIWA interoperability platform. When they are developed in WS-PGRADE, it is possible to use a very useful feature of WS-PGRADE and define almost out of the box, an so called "End User View" that shields the users from the unnecessary implementation details. WS-PGRADE meta-workflows that orchestrate TAVERNA workflows leave to the developers the possibility of customizing the workflows using their language of choice (TAVERNA) while it allows the easy creation of user interfaces through WS-PGRADE that can be used by domain experts that do not want to be concerned with the technical details of the implementation.
- **Iterative Science Workflows** that investigate Science Cases on large multiple data sets. They are best developed in WS-PGRADE to avail themselves of the powerful parameter-sweep capabilities of WS-PGRADE. They invoke sub Science workflows directly in WS-PGRADE and TAVERNA through the SHIWA interoperability platform. At the moment we have not ported yet any of those in the SHIWA Repository as they are still under testing. Although these workflows are not part of this deliverable, every effort will be made to port them by the end of the project or, at the latest, by the end of 2014.

10.3 Science Cases

Name	Description	Workflow IDs
Type II CMEs	Investigates the relationship between shock waves and radio emissions in CMES	5785
Type III Radio Bursts	Distinguishes Coronal Mass Ejections (CMEs) from Co-Rotating Interactive Regions (CIRs)	5787
CME and CIR distinction	Investigates the relationship between radio and "in situ" data for the shock waves generated by Coronal Mass Ejections.	5782

Table 11, Science Cases for the Heliophysics Community

These three science cases have been chosen as they share the common feature of event investigation and they rely on service orchestration rather than big data crunching. This choice highlights the re-usability of their components.

10.3.1 Type II CME Science Case

Eruptive activity in the solar atmosphere can result in the ejection of plasma at over 2000 km/s, known as a coronal mass ejection. Coronal Mass Ejections (CMEs) often drive shock waves through the solar system which can cause intense radio emission (known as a type II) and the eventual detection of the shock at Earth by satellites in-situ. However, although shocks may be readily identified using both radio and in-situ data, few studies have been performed regarding the relationship between these observables.

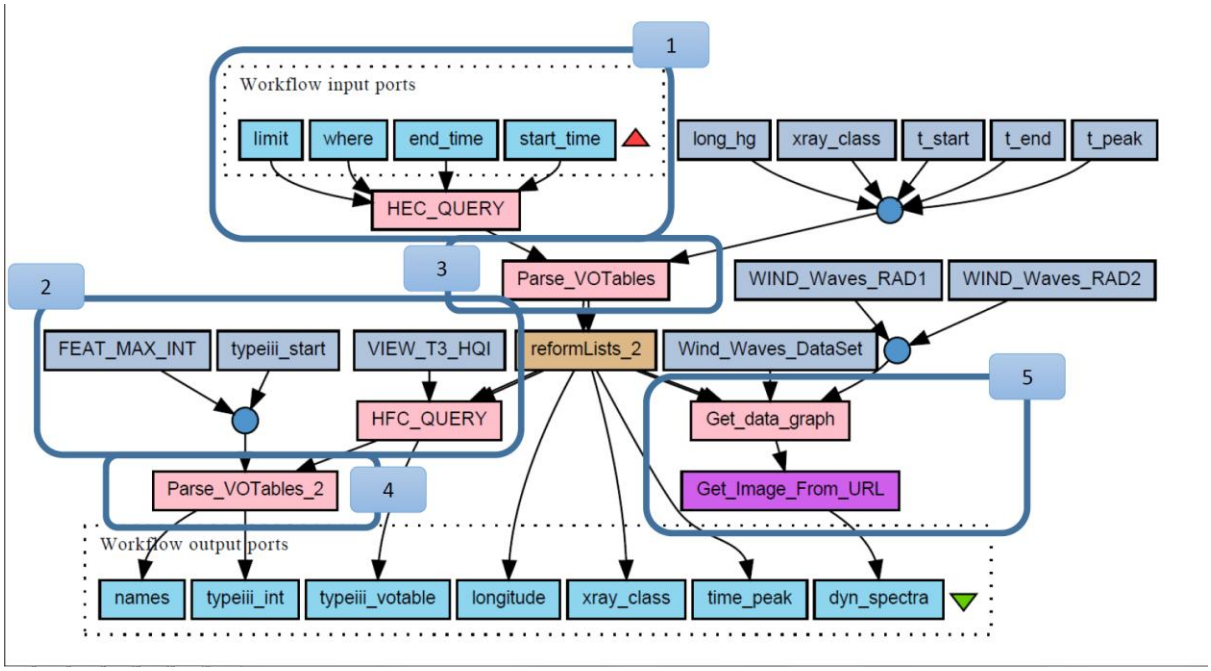


Figure 21, Type II CME Implementation with a TAVERNA meta-workflow

10.3.1.1 Scientific Merit

By relating the radio and in-situ manifestations we may improve our understanding and forecasting of a shock wave impact at Earth.

10.3.1.2 Steps

1. We initially query the wind typeii_soho_cme catalogue for shock radio burst parameters such as start time, end time and observed frequency.
2. The time of the event is then passed to a query of the soho_lasco_cme catalogue, from which a number of CME parameters are chosen, including position angle, angular width, final velocity and velocity uncertainty. The time of the event is also passed to a component that queries the goes_sxr_flare catalogue and obtains the latitude, longitude and strength of the associated solar X-ray flare.
3. This information parsed from the catalogues is then used to run a ballistic propagation model known as SHEBA to produce an ETA at Earth. The ETA of this second iteration is then compared to a catalogue of in-situ shocks and CME detections in order to confirm if there was any positive shock.

10.3.1.3 Example

We analyse an eruptive event that occurred on 11 April 2004. The event was associated with a X-ray flare and a CME which was observed by the LASCO telescopes at 04:30 UT with central position angle 203_, angular width 314, and speed 1645 km s-1.

1. During this time a type II radio burst was observed by WIND/WAVES at 04:20 UT. The associated CME is propagated forward using the SHEBA propagation model to find an initial ETA range of 2004/04/12 05:57:54 21:10:41 UT.
2. We evaluate the solar wind speed at this time from the catalogue of the ACE spacecraft (which measures solar wind speed at Earth). We obtain a solar wind speed of 442 km s□1, which is used to re-evaluate the CME speed and then run SHEBA a second time. This is done to take into account the fact that CMEs are slowed down by a solar wind drag effect.
3. The new ETA range is 2004/04/12 05:57 2004/04/15 03:47 UT . This range is then searched for a positive detection of a shock in-situ. A search for a shock detection by

the SOHO CELIAS instrument reveals a positive shock detection at 2004/04/12 17:35 UT, within the expected time window.

With regard to our science case stated above, the workflow successfully related a radio and in-situ shock detection via a CME propagation model.

10.3.1.4 Related Publications

"*Metaworkflows and Workflow Interoperability for Heliophysics*". Dr. Gabriele Pierantoni, Dr. Eoin Carley, International Workshop on Science Gateways (IWSG 2014), Dublin, June 2014

10.3.1.5 Contacts

- Name(s): Dr. Gabriele Pierantoni
- Email(s): pierang@cs.tcd.ie

10.3.2 CME-CIR Science Case

The Sun may produce a variety of dynamical phenomena that can be observed to propagate far into the Heliosphere. Among these are powerful expulsions of magnetized plasma from the solar atmosphere into the solar system, known as CME. If directed towards Earth they may impact the geomagnetic field, causing a geomagnetic storm, which is particularly dangerous for electrical power grid systems.

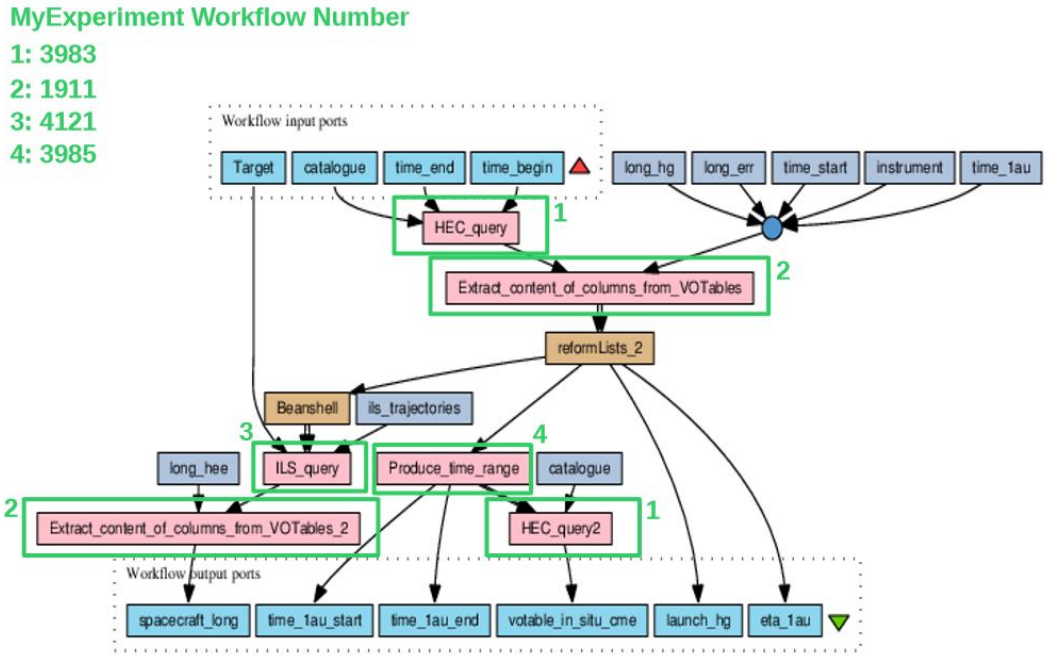


Figure 22, CME-CIR Implementation with a TAVERNA metaworkflow

Another phenomenon which may also cause such a storm (albeit a smaller one than a CME) is a co-rotating interaction region (CIR). CIRs are regions of the solar system where high-speed solar wind crashes into slow speed solar wind, creating a steady shock wave that sweeps around the solar system at the rotation period of the Sun (~30 day rotation period). Since both CMEs and CIRs are potentially damaging, there is an effort to identify and distinguish these phenomena from observational imaging data, mainly from the Heliospheric Imagers (HIs) of the STEREO spacecraft. However, the two phenomena are not easily distinguished in HI and it is difficult to forecast if a CME or CIR will impact Earth. The goal of this workflow is to try and distinguish the characteristics of CMEs and CIRs in imaging

observations by using related in-situ observations (in which the two phenomena *are* easily distinguishable).

10.3.2.1 Scientific Merit

In summary, we use in-situ to identify if a CME or CIR is observed; we then use this to detail the characteristics of each phenomena in the imaging observations. Knowing the characteristics of each in the images should aid future imaging identifications of either phenomena.

10.3.2.2 Steps

Our workflow is outlined as follows:

1. The initial query is to the *stereo_hi_sw_transient* list in a particular time range. From this catalogue we extract the longitude of the source of the transient, the longitude error, the start time (launch time of the transient) and the estimated time of arrival. The time of the event is then passed to the HELIO Instrument Location Service (ILS) to extract the Heliocentric Earth Equatorial longitudes of all available instruments.
2. If the longitude of the spacecraft matches the launch longitude of the solar wind transient, the spacecraft is checked for the possible in-situ detection of a CME or CIR (all available in-situ catalogues above are checked).
3. If a CME or CIR is detected in-situ, both the in-situ data and the imaging data are saved for manual analysis. The procedure is repeated for all detections in the *stereo_hi_sw_transient* list.

10.3.2.3 Example

We firstly searched the *stereo_hi_sw_transient* list in the time range of 2009-05-29 00:00:00 - 2009-05-30 00:00:00 UT (we chose this time range with *a priori* knowledge that a solar wind transient was identified in STEREO HIs). In this time range, a transient was observed in both STEREO ahead and behind spacecraft, traveling at 309 km/s and 300 km/s, respectively. The *stereo_hi_sw_transient* catalogue defines an estimated time of arrival (ETA) based on kinematic modelling of the transient observed in each spacecraft. The ETA based on the observations from STEREO A and B are 2009-06-03 06:55:00 UT and 2009-06-03 21:21:00 UT, respectively.

In order to check the possibility of and in-situ detection of this transient, we compare the estimated propagation longitude of the transient and the spacecraft longitude. The transient launch longitude was -7.8 degrees (Stonyhurst Heliographic, with the central meridian on the Sun-Earth line). The longitude of STEREO A was obtained using the HELIO Instrument Location Service (ILS); it was found to be -61 degrees. Given the longitudinal expanse of CMEs at interplanetary distances (>100 degrees), there is a strong possibility of this transient impacting STEREO A (if indeed the transient was a CME). We chose the ETA and STEREO A define a time window of 48 hrs over which to check for an in-situ detection i.e., check for an impact of STEREO A between 2009-06-02T16:55:00 - 2009-06-04T16:55:00 UT. During this time, the *stereo_impactplastic_icme* catalogue (list of in-situ detections from STEREO) reported a positive detection of a CME at 2009-06-03T16:42 UT. The actual arrival time falls with the ETA window. Hence there is a strong possibility that the solar wind transient observed in images (HI) is the same feature observed by the IMPACT instrument on-board STEREO. In this instance, the workflow is successful in the identification of the solar wind transient observed in HI as a CME, not a CIR.

10.3.2.4 References

- List of STEREO HI transient:

http://hec.helio-vo.eu/hec/hec_gui_free.php?sql=SELECT+*+FROM+hec_catalogue+ORDER+BY+cat_id%3B%0D%0A

- HELIO instrument location service: http://www.helio-vo.eu/services/interfaces/helio-ils_uix.php

10.3.2.5 Related Publications

"*Metaworkflows and Workflow Interoperability for Heliophysics*". Dr. Gabriele Pierantoni, Dr. Eoin Carley, IWSG14, June 2014

10.3.2.6 Contacts

- Name(s): Dr. Gabriele Pierantoni, Dr. Eoin Carley
- Email(s): pierang@cs.tcd.ie

10.3.3 CME-RadioBursts Science Case

Eruptive activity in the solar atmosphere can result in the ejection of plasma known as a coronal mass ejection (CMEs). CMEs often drive shock waves through the solar system, which can cause intense radio emission known as a type II CMEs. However, although shocks may be readily identified using both radio and in-situ data, few studies have been performed regarding the relationship between these observables.

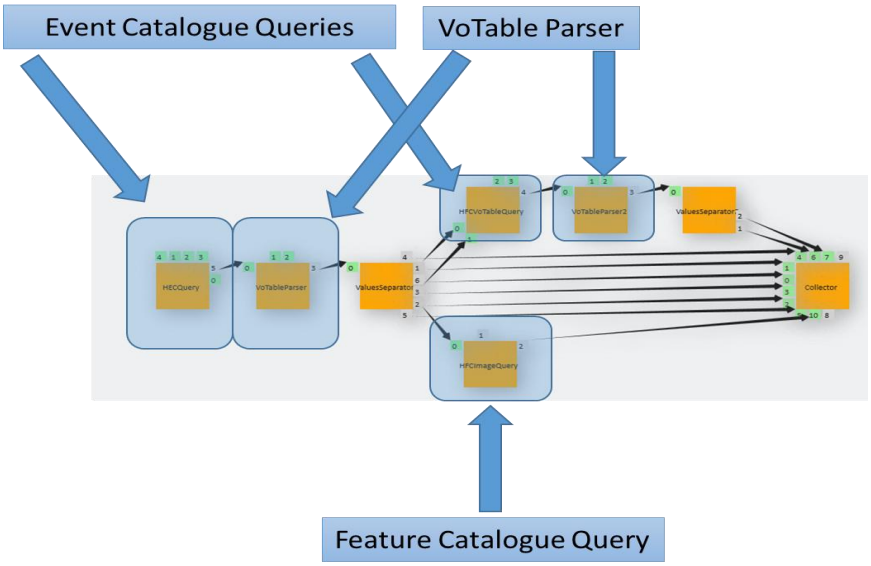


Figure 23, CME Radio Bursts implementation with a WS-PGRADE meta-workflow

10.3.3.1 Scientific Merit

By relating the radio and in-situ manifestations, we may improve our understanding and forecasting of a shock wave impact at Earth.

10.3.3.2 Steps

Our workflow is outlined as follows:

1. We initially query the wind soho_cme_catalogue for shock radio burst parameters such as start time, end time and observed frequency.
2. The time of the event is then passed to a query of the soho_lasco_cme catalogue, from which a number of CME parameters are chosen, including position angle, angular width, final velocity and velocity uncertainty. The time of the event is also passed to a

component that queries the goes_sxr_flare catalogue and obtains the latitude, longitude and strength of the associated solar X-ray flare.

3. This information parsed from the catalogues is then used to run a ballistic propagation model known as SHEBA to produce an Expected Time of Arrival (ETA) at Earth. The ETA of this second iteration is then compared to a catalogue of in-situ shocks and CME detections in order to confirm if there was any positive shock.

10.3.3.3 Example

We analyse an eruptive event that occurred on 11 April 2004. The event was associated with an X-ray flare and a CME which was observed by the LASCO telescopes at 04:30 UT with central position angle 203_, angular width 314, and speed 1645 km s-1.

1. During this time a type II radio burst was observed by WIND/WAVES at 04:20 UT. The associated CME is propagated forward using the SHEBA propagation model to find an initial ETA range of 2004/04/12 05:57:54 21:10:41 UT.
2. We evaluate the solar wind speed at this time from the catalogue of the ACE spacecraft (which measures solar wind speed at Earth). We obtain a solar wind speed of 442 km s□1, which is used to re-evaluate the CME speed and then run SHEBA a second time. This is done to take into account the fact that CMEs are slowed down by a solar wind drag effect.
3. The new ETA range is 2004/04/12 05:57 2004/04/15 03:47 UT. This range is then searched for a positive detection of a shock in-situ. A search for a shock detection by the SOHO CELIAS instrument reveals a positive shock detection at 2004/04/12 17:35 UT, within the expected time window.

With regard to our science case stated above, the workflow successfully related a radio and in-situ shock detection via a CME propagation model.

10.3.3.4 References

- soho_lasco_CME catalogue: http://cdaw.gsfc.nasa.gov/CME_list/
- goes_sxr_flare catalogue: http://sxi.ngdc.noaa.gov/sxi/servlet/sxibrowse
- SHEBA propagation model: https://github.com/dpshelio/SHEBA/blob/master/documentation/sheba.org
- ACE Spacecraft: http://www.srl.caltech.edu/ACE/
- WIND/WAVES Spacecraft: http://www-lep.gsfc.nasa.gov/waves/data_products.html
- Coronal Mass Ejections (CMEs): http://solarscience.msfc.nasa.gov/CMEs.shtml
- X-Ray flares: http://spaceweather.com/glossary/flareclasses.html
- UT: http://spaceweather.com/glossary/flareclasses.html

10.3.3.5 Related Publications

. "*Metaworkflows and Workflow Interoperability for Heliophysics*". Dr. Gabriele Pierantoni, Dr Eoin Carley, IWSG14, June 2014

10.3.3.6 Contacts

- Name(s): Dr. Gabriele Pierantoni, Dr, Eoin Carley
- Email(s): pierang@cs.tcd.ie

10.4 Applications and Workflows

As introduced in paragraphs 10.2 and 10.3, the workflows developed during Year 2 of the project are layered in Science Case meta-workflows and building block sub-workflows. This paragraph gives a brief overview of their nature and goal while technical details can be found in Annex C.

Name	Workflow Engine	Middleware	Sub-Workflows
Type II CMEs	TAVERNA workflows, TAVERNA meta-workflow	Cluster & Web Services	5773, 5775, 5772, 5786
Type III Radio Bursts	WS-PGRADE meta-workflow WS-PGRADE iterative workflow	Cluster & Web Services	5773, 5772, 5778
CME and CIR distinction	TAVERNA workflows, TAVERNA meta-workflow	Cluster & Web Services	5741, 5784, 5783

Table 12, Heliophysics Science Cases and their implementation in meta-workflows

The building blocks developed can be roughly grouped in:

- Query workflows: that invoke web-service based catalogues to retrieve meta-data on features, events and instruments.
- Propagation Model workflows: that invoke the SHEBA propagation model to simulate the propagation of Coronal Mass Ejections (CME), Co-rotating Interactive Regions (CIR) and Solar Energetic Particles (SEP)
- Generic tools: that mainly manipulate VoTables (http://www.ivoa.net/documents/VOTable/) a data and meta-data xml-based standard used in Astrophysics and Heliophysics.

Name	Engine	Type	Middle ware	Description	ID
Type II CMEs	TAVERNA	Meta-workflow (Science Case)	Cluster & Web Service	Science Case Type II CMEs	5785
Type III Radio Bursts	WS-PGRADE	Meta-workflow (Science Case)	Cluster & Web Service	Science Case Type III Radio Bursts	5787
CME and CIR distinction	TAVERNA	Meta-workflow (Science Case)	Cluster & Web Service	Science Case CME and CIR	5782
Remote CME Forward Propagation	TAVERNA	Workflow (Propagation Model)	Web Service	Invokes remotely the Forward Propagation Model for CMEs	5767
Remote CME Backward Propagation	TAVERNA	Workflow (Propagation Model)	Web Service	Invokes remotely (web-service) the Backward Propagation Model for CMEs	5768
Remote CIR Forward Propagation	TAVERNA	Workflow (Propagation Model)	Web Service	Invokes remotely (web-service) the Forward Propagation Model for CIRs	5769
Remote CIR Backward Propagation	TAVERNA	Workflow (Propagation Model)	Web Service	Invokes remotely (web-service) the Backward Propagation Model for CIRs	5760

Remote SEP Forward Propagation	TAVERNA	Workflow (Propagation Model)	Web Service	Invokes remotely (web-service) the Forward Propagation Model for SEPs	5761
Remote SEP Backward Propagation	TAVERNA	Workflow (Propagation Model)	Web Service	Invokes remotely (web-service) the Backward Propagation Model for SEPs	5762
Local CME Forward Propagation	WS-PGRADE	Workflow (Propagation Model)	Cluster	Invokes on a local infrastructure the Forward Propagation Model for CMEs	5765
Local CME Backward Propagation	WS-PGRADE	Workflow (Propagation Model)	Cluster	Invokes on a local infrastructure the Backward Propagation Model for CMEs	5764
Local CIR Forward Propagation	WS-PGRADE	Workflow (Propagation Model)	Cluster	Invokes on a local infrastructure the Forward Propagation Model for CIRs	5767
Local CIR Backward Propagation	WS-PGRADE	Workflow (Propagation Model)	Cluster	Invokes on a local infrastructure the Backward Propagation Model for CIRs	5766
Local SEP Forward Propagation	WS-PGRADE	Workflow (Propagation Model)	Cluster	Invokes on a local infrastructure the Forward Propagation Model for SEPs	5769
Local SEP Backward Propagation	WS-PGRADE	Workflow (Propagation Model)	Cluster	Invokes on a local infrastructure the Backward Propagation Model for SEPs	5768
HECQueryWithParam	TAVERNA	Workflow (Query)	Web Service	Generic MySQL query of the HELIO Event Catalogue	5773
HECQueryOrdered	TAVERNA	Workflow (Query)	Web Service	Ordered query of the HELIO Event Catalogue	5774
HECQuerySimple	TAVERNA	Workflow (Query)	Web Service	Simple query of the HELIO Event Catalogue	5775
HECQueryWSPGRADE	WS-PGRADE	Workflow (Query)	Cluster & Web Service	Simple query of the HELIO Event Catalogue	5741
HFCQuery	TAVERNA	Workflow (Query)	Web Service	Simple query of the HELIO	5776

				Feature Catalogue	
HFCVoTable Query	WS-PGRADE	Workflow (Query)	Cluster & Web Service	Simple query of the HELIO Feature Catalogue	5784
HFCVoImage eQuery	WS-PGRADE	Workflow (Query)	Cluster & Web Service	Image query of the HELIO Feature Catalogue	5783
ICSQuery	TAVERNA	Workflow (Query)	Web Service	Query of the Instrument Capability Service	5777
ILSQuery	TAVERNA	Workflow (Query)	Web Service	Query of the Instrument Location Service	5778
WindWaves Quicklook	TAVERNA	Workflow (Query)	Web Service	Returns quicklook from the Wind Waves instrument	5786
VoTableExtractor	TAVERNA	Workflow(Tool)	Web Service	Parser VoTables	5772

Table 13, Workflows ported in Year 2

10.5 Applications Usage

The user scenarios of the Heliophysics community cover different usage patterns of both workflows and meta-workflows and different users.

10.5.1 TAVERNA Workflows and Meta-workflows

Developers and users can design and execute TAVERNA workflows and meta-workflows through the TAVERNA workbench, whereas users can store and retrieve workflows from the myExperiment web site that stores all the TAVERNA workflows

10.5.2 WS-PGRADE Workflows and Meta-workflows

Developers and users that deal with pure WS-PGRADE workflows have a variety of choices to design and execute the workflows. They can design abstract and concrete workflows from both the HELIOGate Portal and the SHIWA Simulation Platform. Users that do not want to be bothered by technical details can use the “simple user views” that expose only information on the parameters and the status of the execution. The SHIWA Repository containing all workflows and meta-workflows (TAVERNA, WS-PGRADE) can be accessed by both the SHIWA Simulation Platform and the HELIOGate portal.

10.5.3 Workflows Interoperability

WS-PGRADE meta-workflows can be executed with the same interface as simple WS-PGRADE workflows from the SHIWA Simulation Platform, interoperability execution from HELIOGate was supported by the new SHIWA Submission Service.,

11 Life Sciences

Life sciences (LS) comprise the fields of science that involve the scientific study of living organisms, such as microorganisms, plants, animals, and human beings. The Life Sciences community is represented in ER-Flow by the Academic Medical Centre of the University of Amsterdam. This community focuses on biomedical research, which is a subfield of life sciences with the aim of better understanding the mechanisms of diseases, how they manifest themselves in detectable ways, and how they can be influenced to treat the patient. The final goal of biomedical research is to improve healthcare with better diagnostics, prognosis, and treatment by means of interventions with drugs, therapy of various types (e.g., radiotherapy), surgery, or changes in life style. Moreover, better understanding of diseases can help disease prevention and general improvement of health and well-being in the society.

The e-science group of the AMC participates in ER-Flow. The members of this group communicate with diverse biomedical researchers at AMC, including the following research domains: neurosciences, next generation sequencing, biostatistics, mass spectrometry and molecular docking. These researchers cover a large spectrum of expertise and profiles, including researchers or domain scientists that run workflows prepared by others developers of new data analysis methods (e.g., medical imaging or bioinformatics) who typically build and run their own workflows, and e-Science researchers, who port applications to the e-infrastructures in collaboration with domain scientists, and also develop and maintain science gateways for these biomedical researchers. For the purposes in ER-Flow, the developers and e-science researchers form one group, which have been responsible for porting the applications described in this deliverable.

In the second project year, we approached some Life Sciences communities external to the scope of the ER-Flow project, with the goal of promoting scientific workflow technologies and workflow interoperability. As a result of this effort two new groups have been successfully involved in the ER-flow project. The first is the German SOMNO.NETZ project (https://www.somnonetz.de/) lead by the University of Applied Sciences Berlin. This collaboration resulted in three WS-PGRADE workflows for cloud infrastructure, which are also covered in this deliverable. The second group is composed by bioinformaticians who are heavy users of Galaxy and TAVERNA workflow management systems, respectively at the Leiden University Medical Center and the Leiden University, NL. In collaboration we developed an application for analysis of RNA-seq datasets. This collaboration resulted in a tutorial presented at a major European conference in bioinformatics and various workflows and meta-workflows. At the time of the writing a large experiment is being carried out using these workflows. See also section 11.3.3. In spite of these successful collaborations, we realize that the life science community represented in ER-flow is still limited.

The applications ported by the Life Sciences community to the SHIWA platform in the second project year can be organized into three major scientific application areas, namely neuroscience, genomics and drug development. The science cases reported in this document cover all these domains, namely, study of brain connectivity with Diffusion Tensor Imaging (DTI), brain scans segmentation from structural Magnetic Resonance Imaging (MRI) scans, DNA and RNA sequencing to identify markers of disease, and virtual drug compound screening. The workflows implementing the necessary applications have been implemented from scratch or adapted from other workflows developed in Y1 by our own community or developed by others and available in the SHIWA repository. Finally, some workflows were created as a result of collaboration with the SOMNO.NETZ project based on the MoU signed during this project period.

11.6 Technical Background

The applications are characterized by large input files, and produce a large number of output files, which are normally combined into a single archive. Some perform “data reduction” (input size larger then output size), whereas others perform “data production” (output size smaller than input size). The applications have various workflow implementations, both for MOTEUR and for WS-PGRADE workflow systems, which are currently adopted at the AMC as back-ends for the science gateways. The workflow implementations of these applications are defined as templates of processing chains that can be applied to a single unit or to a list of input values or files. Currently, all the code used inside workflow tasks is itself sequential. Parallelism is obtained by distributing the data (or parts of the data) to processes that are started in parallel by the workflow systems on different data and parameter settings.

The distributed infrastructure used to run AMC applications is the Dutch grid infrastructure, which is part of EGI. The middleware used is gLite and follows the regular EMI releases as recommended by EGI operations. The AMC operates its own virtual organization (VLEMED) since 2005. The resources available for the VLEMED VO are distributed among 14 sites in The Netherlands, some of which are part of the Life Science Grid (LSG). The VLEMED VO has access to compute elements, storage elements (SRM), an LFC file catalogue, and various other gLite generic services for proxy, job and information management. Grid resources are coordinated by workflows enacted from both MOTEUR and WS-PGRADE engines. Some of the workflows have been additionally ported to a local cluster (PBS) using WS-PGRADE. This cluster is located in the demilitarized zone (DMZ) of the AMC network, offering a more reserved environment for privacy-sensitive applications.

The SOMNO.NETZ community employs a different infrastructure. The data is stored in a storage server called XNAT, and it is moved to a private cloud for processing. This cloud uses Openstack as its middleware. Traditionally, a process is started via the XNAT system, and the job, when on the worker node on the cloud, reads the data directly from XNAT and in the end write the output back to XNAT. In the experiments performed during ER-Flow and SOMNO.NETZ collaboration, WS-PGRADE workflows were developed that would be started from the WS-PGRADE interface.

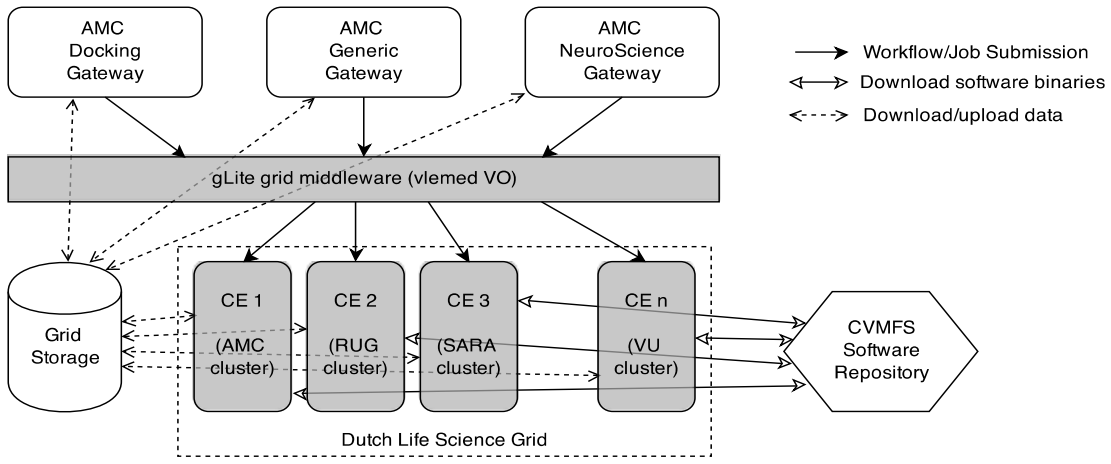


Figure 24, Technical Background of the Life Science Community

We have negotiated with the Grid support and maintenance team to deploy the Cern VM File System (CVMFS) for software distribution and versioning management. This entails maintaining a central software repository on a server, plus installing CVMFS clients on the grid worker nodes. With the help of CVMFS, the workflows do not need to carry around or download the software binaries anymore. Instead, the software is loaded using the ‘module’ tools, and a caching system helps reduce the network traffic for common and large software files. This has improved greatly the performance of our workflows on the grid. On the other

hand, this makes the future shift to the cloud much easier, because our lightweight workflows can now run also on the cloud and fetch the required software on demand.

11.7 Workflow Usage in the Community

There are roughly two types of users in the Life Science community represented in ER-flow: those who directly develop (workflow developers) and run the workflows for a particular scientific study, and those who only run existing workflows from a customized interface (domain scientists). The first group uses three generic WS-PGRADE portals: the SSP provided by ER-flow, the AMC portal provided by the SCI-BUS project, and the SZTAKI portal. The last one is occasionally used for testing advanced releases that are not yet implemented in the other two portals. The second group uses the AMC customized science gateways developed in SCI-BUS: for neuroscience (https://neuro.ebioscience.amc.nl) and virtual screening (https://docking.ebioscience.amc.nl). These gateways are developed and maintained with funding from SCI-BUS and COMMIT/ projects as well as the University of Amsterdam HPC Fund.

The workflow development strategy is the same in both cases. A workflow is developed and tested by a workflow developer, added to the repository, and executed in the scope of some scientific experiment. Then, the workflow is executed repeatedly as part of various scientific experiments on different data sets or parameters. Each experiment in turn requires hundreds to thousands of inputs to be processed (depending on the application domain). Normally, this is achieved by applying “parameter sweep” functionality of WS-PGRADE. This is useful for cases in neuroscience domains like DTI processing and brain segmentation, in which data is usually a few hundred MB for each run and one experiment processes hundreds of brain images. Alternatively, some workflows take the whole input at once and split the computation internally. This is adopted for example in docking experiments, where the size of the input is small, but thousands to millions of inputs may be processed in one experiment.

Note, however, that since the collection of data for a new experiment may take months or even years, the workflows might not be executed very often or many times. An exception is the Autodock Vina application, which is a simulation tool that can be executed many times to test various hypotheses. Even in this case the number of workflow executions can be quite limited because the virtual screening phase of the experiment is short; after the simulation, wet-lab experiments are needed to follow up on the hints and advance the research, generating new hypotheses, etc. Validation and iteration can take years. Although this seems to be a slow process, the impact of the findings can be huge, because they directly point to possible improvements in care and treatment.

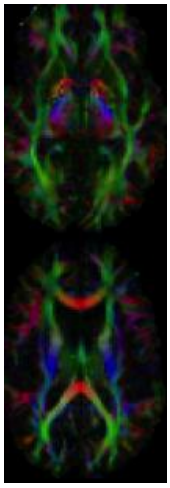
11.8 Science Cases

The science cases presented here illustrate scientific experiments that have been performed based on data generated or analysed using the workflows developed in Year 1 and Year 2. Table 15 summarizes the information that will be detailed in the subsequent sections.

Name	Description	Workflow IDs
DTI	Diffusion Tensor Imaging helps understand the brain	4250, 5709, 5665, 4858, 4601, 5145, 5711
Brain segmentation	Structural brain imaging to helps discovery of disease markers	4251, 5704. 5951, 5952, 5953, 5954, 5955
NGS	Identifying genes associated with disease	4603, 4604, 4602
RNAseq	RNA sequencing for identifying candidate genes	5905, 5906, 5907
Drug screening	Virtual drug screening for controlling atherosclerotic plaque	4605, 5658

Table 14. Science Cases of the Life Science Community

11.8.1 DTI Science Case: Diffusion Tensor Imaging helps understand brain connectivity



Diffusion tensor imaging (DTI) is a magnetic resonance imaging (MRI) modality that enables in-vivo study of brain structure and function. It helps understand the integrity and shape of the white matter fiber tracts that connect different brain regions. The analysis of DTI data requires various specialized and complex steps using a large number of tools. All the necessary steps have been encapsulated into a workflow to facilitate data analysis for the neuroscientists. This workflow is offered to the users from a science gateway with intuitive interface, and it runs on the Dutch grid infrastructure.

11.8.1.1 Scientific Merit

Brain function and structure is fundamental to understand changes in the brain due to disease or other (external) factors. DTI data is widely used by neuroscientists and clinical researchers in studies that attempt to correlate patient symptoms to measurable modifications in brain function and structure. Much on-going research in neurosciences today aims at establishing DTI- and other MRI-based biomarkers of brain disease.

11.8.1.2 Steps

The analysis of DTI data involves various steps from data pre-processing, calculation of diffusion properties such as FA (fractional anisotropy), and finally to determination of white matter fiber bundles with some kind of fiber tracking algorithm.

In the example presented below the first two steps were executed as a workflow. All steps are encapsulated in one single job that is executed in a parameter sweep workflow for each DTI dataset.

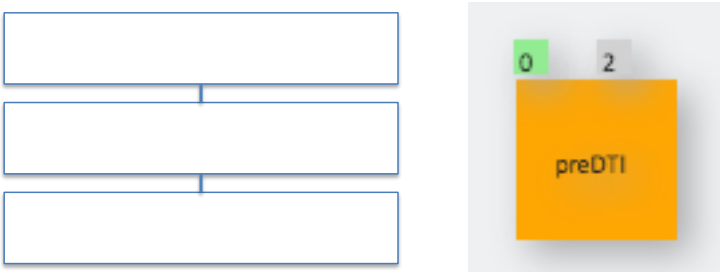


Figure 25, Steps for DTI data processing and feature calculation and the WS-PGrade workflow

11.8.1.3 Example

Note: The text below is based on an article from EGI Case Studies:
http://www.egi.eu/case-studies/medical/combat_stress.html
It illustrates the most prestigious study done so far based on data analysed with the DTI workflows.

Veterans returning from active duty endure many challenges to readapt to a civilian life. For some there is the added complication of combat stress, which affects their memory, attention span and other cognitive functions. But how? And for how long?

Neuroscientist Guido van Wingen and colleagues at the Academic Medical Centre (AMC) in Amsterdam monitored a group of soldiers from before their first deployment to Afghanistan until 18 months after their return to civilian life. The idea was to look how combat stress affects brain areas supporting cognitive functions such as memory and attention. The team

used the grid-enabled e-bioinfra science gateway to process and analyse 118 brain scans from 33 soldiers and 26 civilians used as controls. Thanks to a friendly user interface and the computing power of the grid, a workload of several weeks was condensed into two days.

The conclusions, published in the *Proceedings of the National Academy of Sciences*, show that combat stress impairs cognition by affecting the midbrain and its link with the prefrontal cortex, and that this is largely reversible but could have an impact on future social and cognitive functions.

11.8.1.4 Related Publications

These publications have been produced based on data analysis realized with the execution of workflows for DTI data analysis on the Dutch grid infrastructure.

1. Guido A. van Wingen, Elbert Geuze, Matthan W.A. Caan, Tamás Kozicz, Silvia D. Olabarriaga, Damiaan Denys, Eric Vermetten, Guillén Fernández (2012). Persistent and reversible consequences of combat stress on the mesofrontal circuit and cognition. Proceedings of the National Academy of Sciences of the USA. PNAS September 18, 2012 vol. 109 no. 38 pp. 15508-15513
2. B. de Kwaasteniet, E. Ruhe, M. Caan, M. Rive, S. Olabarriaga, M. Groefsema, L. Heesink, G. van Wingen, and D. Denys, Relation between structural and functional connectivity in major depressivedis- order Biological Psychiatry, no. 0, 2013.
3. Bart D. Peters, M.D., Marise Machielsen, Wendela Hoen, M.D., Matthan W. Caan, Philip R. Szeszko, Anil K. Malhotra, Silvia D. Olabarriaga, Lieuwe de Haan. Polyunsaturated Fatty Acid Concentration Predicts Myelin Integrity in Early Psychosis.. Schizophrenia Bulletin, Schizophr Bull. 2012 Aug 27 (epub ahead)

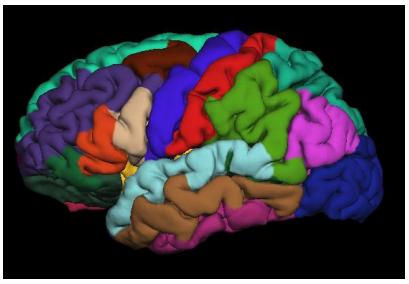
The neuroscience gateway where the DTI application is available is described in the following publication:

Shayan Shahand, Ammar Benabdelkader, Mahdi Jaghouri, Jordi Huguet, Mostapha Al Mourabit, Matthan Caan, Antoine van Kampen and Silvia Olabarriaga. A Data-Centric Science Gateway for Computational Neuroscience. Concurrency and Computation: Practice and Experience, epub 14 April 2014

11.8.1.5 Contacts

- Silvia Delgado Olabarriaga, S.D.Olabarriaga@amc.nl
- Matthan Caan, M.W.A.Caan@amc.uva.nl

11.8.2 Brain Segmentation Science Case: Structural brain imaging helps discovery of disease markers



Magnetic resonance imaging (MRI) provides in-vivo information about brain structure, which is important to identify locations that may be correlated to disease. Characteristics of a brain region, such as shape and volume, can be associated with a phenotype, and serve as “biomarkers”. Brain segmentation is also used to facilitate the analysis of brain connectivity between regions, in combination with methods described in the DTI Science Case.

The analysis of brain structural data requires a complex set of steps where the individual regions of the brain are identified, delineated and “labelled”, in a process called “brain segmentation”. This is a challenging task: manual segmentation is time-consuming, and automated segmentation is very difficult due to brain variability. The Freesurfer software suite is a tool that has been developed and optimized for automated brain segmentation. It

includes 30+ steps that have been encapsulated into a workflow; the complete execution takes several hours (12 to 48+ hours). This workflow has been ported for the Dutch grid infrastructure, being offered to the users from a science gateway with intuitive interface.

11.8.2.1 Scientific Merit

Brain segmentation is a necessary step in any study that uses structural brain scans. As the size of neuroscience and clinical studies grow, also the challenges to perform brain segmentation are increasing. The automated labelling of brain regions empowers neuroscientists and clinical researchers to more easily analyse the characteristics of individual regions, and to develop markers that can be used to detect disease, follow their progress, and control treatment effectiveness along time.

11.8.2.2 Steps

Brain segmentation is a complex process – the figure below illustrates some of the steps involved.

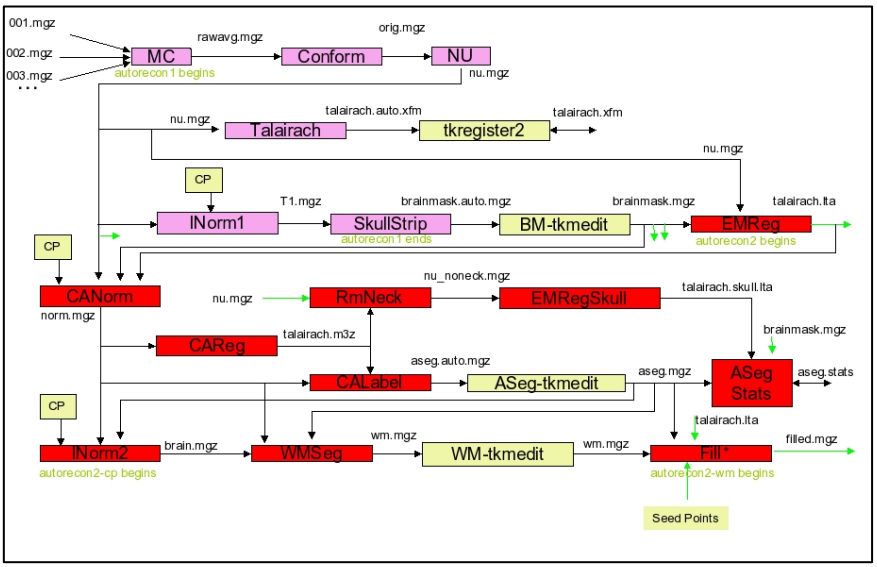


Figure 26, Impression of the Freesurfer brain segmenation analysis pipeline (recon-all option). Extracted from http://surfer.nmr.mgh.harvard.edu/fswiki

All these steps are provided by the Freesurfer toolbox, which can be activated with different execution options to achieve different goals. We implemented various workflows to run Freesurfer with different options. Below you see the workflow that runs the complete brain segmentation pipeline. Freesurfer execution is split into various steps with checkpointing (although inside one job) to improve robustness of the execution, which can take 1-2 days on the resources of the Dutch grid infrastructure available for our Virtual Organization.

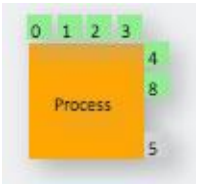


Figure 27, Freesurfer recon-all Workflow

11.8.2.3 Example

Note: here we describe only one example of various on-going studies that are using the brain segmentation workflow.

Patients with mild cognitive impairment do not always develop dementia. In such cases, abnormal neuropsychological test results may not validly reflect cognitive symptoms due to brain disease, and the usual brain–behaviour relationships may be absent. This study examined the associations between hippocampal volume and memory performance to the results of clinical tests adopted to assess the cognitive condition of 170 patients. The results show that the results obtained by psychological tests correlate well to the two other markers in a group of patients, but they do not correlate well in another group of younger patients. These results have raised various questions about the validity of the tests used in clinical practice to assess the stage or develop prognosis of cognitive impairment diseases.

11.8.2.4 Related Publications

1. Rienstra A, Groot PF, Spaan PE, Majoie CB, Nederveen AJ, Walstra GJ, de Jonghe JF, van Gool WA, Olabarriaga SD, Korkhov VV, Schmand B. (2012) Symptom validity testing in memory clinics: Hippocampal-memory associations and relevance for diagnosing mild cognitive impairment. J Clin Exp Neuropsychol. 2012 Dec 11. [Epub ahead of print]
2. Shayan Shahand, Ammar Benabdelkader, Mahdi Jaghouri, Jordi Huguet, Mostapha Al Mourabit, Matthan Caan, Antoine van Kampen and Silvia Olabarriaga. A Data-Centric Science Gateway for Computational Neuroscience. Concurrency and Computation: Practice and Experience, epub 14 April 2014

11.8.2.5 Contacts

- Silvia Delgado Olabarriaga, S.D.Olabarriaga@amc.nl
- Matthan Caan, M.W.A.Caan@amc.uva.nl

11.8.3 NGS Science Case: Identifying genes that correlate to disease

High-throughput experimental methods in Molecular Biology, such as next generation sequencing (NGS), provide the quantitative basis for gaining a better understanding of human disease. However, multifactorial diseases such as diabetes and atherosclerosis are complex disorders involving hundreds of genes and many developmental and environmental factors. Computational methods are needed that can uncover the molecular networks perturbed by disease. The set of WS-PGRADE workflows developed during the ER-flow project provide a toolbox to analyse DNA sequencing data. They execute on the Dutch grid using resources of the VLEMED VO. They can be executed from the generic interface of the SHIWA Simulation Platform and from the AMC generic WS-PGRADE portal.

11.8.3.1 Scientific Merit

NGS is the holy grail of clinical research due to the promise to reveal the most hidden mechanisms of disease. Massive amounts of NGS data are being collected all around the globe. These workflows help the challenging task of extracting information from such big data.

11.8.3.2 Steps

The analysis of NGS data involves a large number of steps that have been implemented in individual workflows, such as alignment of sequences to a reference genome; sequence assembly, identification of insertions/deletions, identification of SNP (single nucleotide polymorphism), etc.

11.8.3.3 Example

Note: this is the most prestigious example of grid workflow usage for genomics that we have in our organization so far. In this case the sequence alignment workflow (BWA, Figure 28)

was executed to validate the results against a larger population, upon request of the reviewers of that journal. The authors chose to validate their findings against the Genome of the Netherlands dataset (750 individuals). The processing needed to be done within the given rebuttal time, therefore the usage of a grid made it possible for the authors to be able to respond within the deadline. The workflow suite developed in ER-flow actually covers much more than sequence alignment.

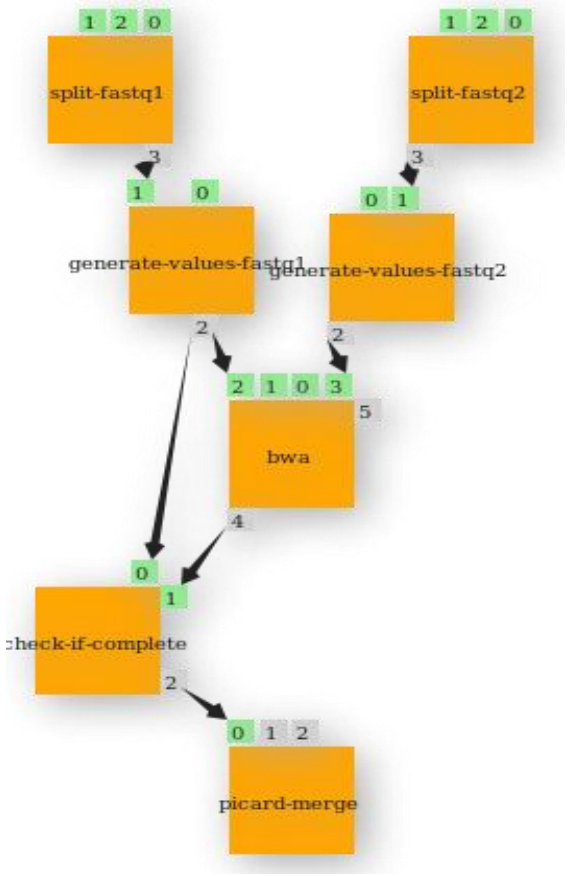


Figure 28, Implementation of BWA workflow for sequence alignment in WS-PGRADE

Nicolaides-Baraitser syndrome (NBS) is characterized by sparse hair, distinctive facial morphology, distal-limb anomalies and intellectual disability. In a study conducted at the AMC the exomes of ten individuals with NBS and identified heterozygous variants in SMARCA2 in eight of them. Exomes are an interesting region of interest in genes corresponding to the part of the genome formed by exons, the sequences which when transcribed remain within the mature RNA after introns are removed by RNA splicing. Exons are related to transcription into proteins and are considered to indicate more direct relationships with disease manifestation.

Extended molecular screening identified non-synonymous SMARCA2 mutations in 36 of 44 individuals with NBS. These mutations were confirmed to be de novo when parental samples were available. SMARCA2 encodes the core catalytic unit of the SWI/SNF ATP-dependent chromatin re-modelling complex that is involved in the regulation of gene transcription. The identification of SMARCA2 mutations in humans provides insight into the function of the Snf2 helicase family.

11.8.3.4 Related Publications

Van Houdt JK, Nowakowska BA, Sousa SB, van Schaik BD, Seuntjens E, Avonce N, Sifrim A, Abdul-Rahman OA, van den Boogaard MJ, Bottani A, Castori M, Cormier-Daire V, Deardorff MA, Filges I, Fryer A, Fryns JP, Gana S, Garavelli L, Gillessen-Kaesbach G,

Hall BD, Horn D, Huylebroeck D, Klapecki J, Krajewska-Walasek M, Kuechler A, Lines MA, Maas S, Macdermot KD, McKee S, Magee A, de Man SA, Moreau Y, Morice-Picard F, Obersztyn E, Pilch J, Rosser E, Shannon N, Stolte-Dijkstra I, Van Dijck P, Vilain C, Vogels A, Wakeling E, Wieczorek D, Wilson L, Zuffardi O, van Kampen AH, Devriendt K, Hennekam R, Vermeesch JR. (2012) Heterozygous missense mutations in SMARCA2 cause Nicolaides-Baraitser syndrome. Nature Genetics, 44(4), 445-9

11.8.3.5 Contacts

- Barbera van Schaik, b.d.vanschaik@amc.uva.nl
- Antoine van Kampen,a.h.vankampen@amc.uva.nl
- Silvia Olabarriaga, S.D.Olabarriaga@amc.uva.nl

11.8.4 RNA-Seq Science Case: NGS helps reveal complex differential genes and transcript expression

RNA-seq (RNA Sequencing) uses the capabilities of next-generation sequencing to reveal a snapshot of RNA presence in the nucleotide of a cell at a particular time. This technique provides the ability to study gene expression, gene fusion, and mutations with more resolution than it was possible using microarrays technology. Samples of two groups with different conditions are compared to determine genetic differences that might explain the observed phenotype differences. For example, samples can be obtained from a group of patients with a given disease, and from another group without that disease, and further combined to generate a list of candidate genes or mutations that might be associated with the disease. Another type of study compares data acquired before and after application of a specific treatment.

11.8.4.1 Scientific Merit

Finding associations between genes, their expression, and the phenotypes is essential to understand the processes of disease, their diagnosis, and a treatment – for example with drugs. The findings from in-vitro experiments need to be confirmed with expensive wet-lab experiments, therefore there is large interest in optimizing the data analysis to provide a short list of highly relevant candidate genes.

11.8.4.2 Steps

The analysis of RNA-seq data involves various steps that can be roughly divided into two parts: discovery of candidate genes and annotation of genes. The first part begins with raw sequencing reads and produces a transcriptome assembly, lists of differentially expressed and regulated genes and transcripts. These steps are implemented with a Galaxy workflow following the TUXEDO pipeline described in [Trapnell et al, 2012] (see figure and publication list below). This pipeline has been ported as a WS-PGRADE workflow to reach higher scalability via execution on a grid infrastructure. The second part, for gene annotation, performed with a TAVERNA workflow, consists of enrichment of metadata about the list of genes generated by the first group using additional databases and ontologies.

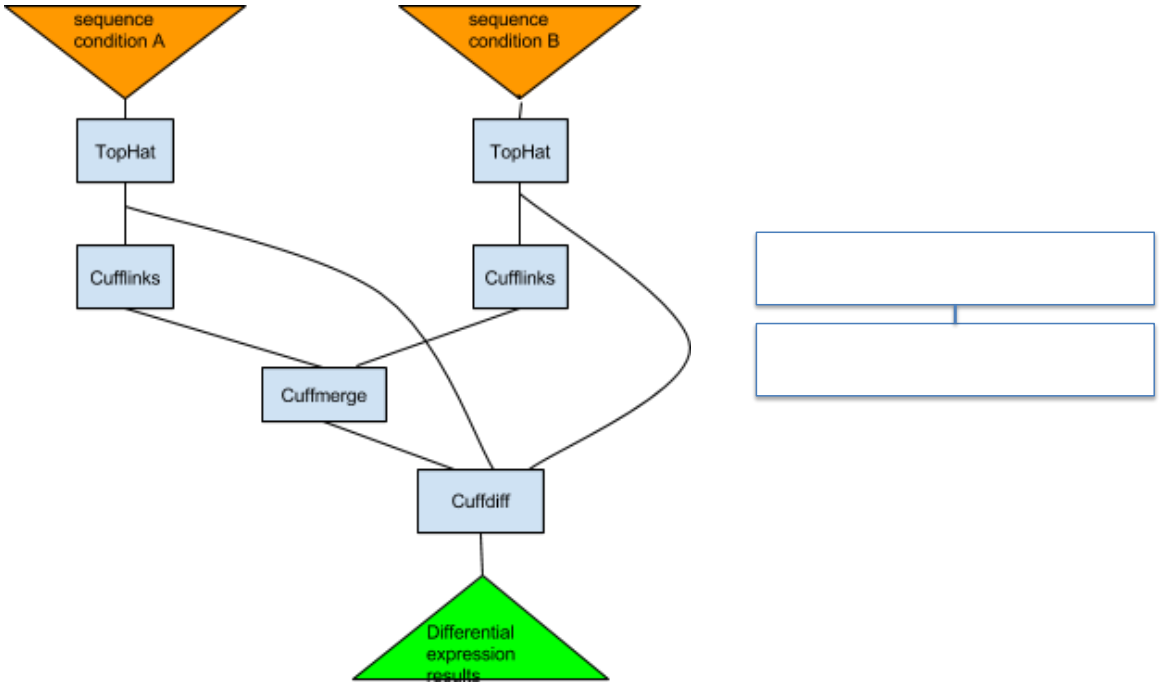


Figure 29, TUXEDO pipeline and complete pipeline for enriched differential gene expression analysis from RNA-seq data

11.8.4.3 Example

RNA-seq experiments must be analyzed with robust, efficient and statistically principled algorithms. Fortunately, the bioinformatics community has been hard at work developing mathematics, statistics and computer science for RNA-seq and building these ideas into software tools. RNA-seq analysis tools generally fall into three categories: (i) those for read alignment; (ii) those for transcript assembly or genome annotation; and (iii) those for transcript and gene quantification. Two popular tools are widely used that together serve all three roles. TopHat (http://tophat. cbcb.umd.edu/) aligns reads to the genome and discovers transcript splice sites. These alignments are used during downstream analysis in several ways. Cufflinks (http://cufflinks.cbcb.umd.edu/) uses this map against the genome to assemble the reads into transcripts. Cuffdiff, a part of the Cufflinks package, takes the aligned reads from two or more conditions and reports genes and transcripts that are differentially expressed using a rigorous statistical analysis. These tools are gaining wide acceptance and have been used in a number of recent high-resolution transcriptome studies.

RNA-seq experiments can serve many purposes, one of the most used cases is a workflow that aims to compare the transcriptome profiles of two or more biological conditions, such as a wild-type versus mutant or control versus knockdown experiments. For simplicity, such experiment can compare only two biological conditions, although the software is designed to support many more, including time-course experiments.

11.8.4.4 Related Publications

Trapnell et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols 7, 562–578 (2012)

11.8.4.5 Contacts

- Silvia Delgado Olabarriaga, S.D.Olabarriaga@amc.nl
- Leon Mei, hailiang.mei@dtls.nl

- Katy Wolstencroft, k.j.wolstencroft@liacs.leidenuniv.nl
- Vladimir Korkhov, vkorkhov@gmail.com

11.8.5 Drug screening use case

In drug discovery research, a drug candidate is typically a small molecule (ligand) that can make a strong binding to the drug target. The target is typically a protein that is either a receptor responsible for transmitting signals to a cell, or an enzyme that enhances the rate of chemical reactions. A drug candidate can interfere with these biological processes incurred by the target molecule, if it can bind with high affinity at specific binding sites. Therefore, scientists usually search available libraries of ligands for putative drug candidates; this process is called screening or docking. Biochemical screening, however, requires expensive lab equipment and takes considerable time. Alternatively, the initial phase of drug discovery nowadays involves virtual screening software tools that simulate binding of ligands and calculate binding affinities. Only drug candidates with high affinities (resulted from simulation) will be further studied in the wet lab.

11.8.5.1 Scientific Merit

AutoDock Vina is one of the most advanced software suites in molecular modelling simulation. It is especially effective for Protein-ligand docking. AutoDock Vina is available under the Apache license. AutoDock Vina achieves an approximately two orders of magnitude speed-up compared to the AutoDock 4 molecular docking software, while also significantly improving the accuracy of the binding mode predictions. Further speed-up is achieved from parallelism, by using multithreading on multi-core machines. AutoDock Vina automatically calculates the grid maps and clusters the grid maps and clusters the results in a way transparent to the user.

11.8.5.2 Steps

The workflow (shown below) starts with the Prepare component, which takes as input, the receptor file (on port 0), the configuration file (on port 1) and a list of ligands as a zipped file (on port 2). The Prepare component splits the ligands into distinct groups and passes them to the AdVina component, along with the receptor and configuration files. Several instances of the AdVina component are executed to process these ligand groups in parallel. Finally, the Collect component collects all the outputs, creates a sorted list of the binding energies, and packs the outputs into a single compressed file.

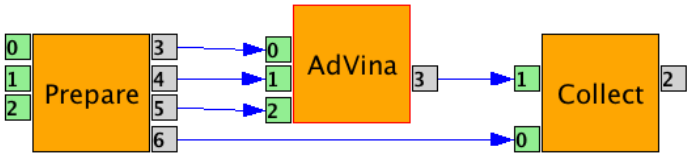


Figure 30, WS-PGrade workflow for parallel execution of AutoDock Vina on the Grid

The workflow dynamically maximizes parallelization granularity based on two restrictions: a minimum number of ligands per job is needed to minimize the overhead versus computation time; and, a maximum splitting factor is used to lower the overhead on gUSE.

11.8.5.3 Example

A particular molecule related to the development of atherosclerosis has been studied for many years at the AMC. This molecule seems to be a natural mechanism to protect blood vessels from plaque development (=slows down the development of plaque).

There is interest in controlling the activation of this molecule with drugs, however this is not easy to do. The general understanding today is that it cannot be artificially controlled

however the researchers of the AMC still have hope. Virtual screening for compounds that could activate this molecule is therefore even more challenging because there are no obvious binding sites to “look” at. Lots of docking simulations are needed. It is like looking for a needle in a haystack – without knowing upfront that the needle is actually there. Large screening experiments have been started at the AMC from the docking gateway to investigate means to activate this molecule. It will take many years still to find out whether this molecule can be activated with drugs.

PS: the name of the molecule is omitted here to protect intellectual rights of the AMC researchers.

11.8.5.4 Related Publications

MM Jaghoori, et al. A Grid-Enabled Virtual Screening Gateway. In the 6th International Workshop on Science Gateways (IWSG 2014), pages 24-29.

11.8.5.5 Contacts

- Silvia Delgado Olabarriaga, S.D.Olabarriaga@amc.nl
- Boris Bleijlevens, B.Bleijlevens@amc.uva.nl
- Carlie J.M. de Vries, c.j.devries@amc.uva.nl

11.9 Applications and Workflows

The workflows developed in Year 2 are summarized below. Further details about the workflows are available in Annex D.

Name	Engine	Type	Middleware	ID
DTI pre-processing (CVMFS)	WS-PGRADE	workflow	gLite	5956
FSL BedpostX (CVMFS)	WS-PGRADE	workflow	gLite	5915
DTI Population Registration	MOTEUR	workflow	gLite	4601
	WS-PGRADE	workflow	gLite	5145
DTI Population Registration (CVMFS)	WS-PGRADE	workflow	gLite	5602
Freesurfer pial	WS-PGRADE	workflow	gLite	5954
Freesurfer param	WS-PGRADE	workflow	gLite	5952
Freesurfer recon-all (CVMFS)	WS-PGRADE	workflow	gLite	5951
Freesurfer pial (CVMFS)	WS-PGRADE	workflow	gLite	5955
Freesurfer param (CVMFS)	WS-PGRADE	workflow	gLite	5953
Trac-all (CVMFS)	WS-PGRADE	workflow	gLite	5958
Autodock Vina with sorting all energies	WS-PGRADE	workflow	gLite	5658
CQRS	WS-PGRADE	workflow	OpenStack	5254
CQRS_DL	WS-PGRADE	workflow	OpenStack	5253
CQRS_ONE	WS-PGRADE	workflow	OpenStack	5255
Tophat	WS-PGRADE	workflow	gLite	5906
Tuxedo	WS-PGRADE	workflow	gLite	5905
RNASeq	WS-PGRADE +TAVERNA	meta-workflow	gLite + Web Services	5907

Table 15, Life Sciences workflows

11.10 Applications Usage

The applications are used to process datasets owned by the biomedical researchers. Normally, the biomedical researcher starts the workflow from a customized interface of a science gateway. In this case, they upload the data to a location from which the files are automatically transported to the grid, and where the results are found after the processing is complete. The science gateway manages the data transport and provides workflow monitoring services.

	Docking gateway	NeuroScience gateway
WfMS	WS-PGRADE	WS-PGRADE
DCI	Dutch Grid	Dutch Grid
url gateway	http://docking.ebioscience.amc.nl/	http://neuro.ebioscience.amc.nl/
Applications (Aug. 2014)	Customized interface for AutoDock Vina	DTI pre-processing, FSL BedpostX, Freesurfer, Trac-all

Table 16, Characteristics of science gateways available to run Life Science applications.

The e-bioscience team of the AMC currently operates two customized science gateways coined “Docking gateway” and “NeuroScience gateway”. The first is intended for

Biochemists at the AMC and allows them to run AutoDock Vina experiments on thousands to millions of inputs. The second gateway covers the DTI processing and Brain Segmentation science cases. Both gateways are based on the platform provided by the SCI-BUS project (gUSE/WS-PGRADE and Liferay). The characteristics of these gateways are summarized in Table 16.

Advanced users also start workflows from the WS-PGRADE workflow developer's web interface, or from command-line interfaces to submit workflows to the MOTEUR web service available at the AMC. In this case, the user him/herself uploads the input data to the grid resource and downloads the results afterwards. The WS-PGRADE generic gateway of the AMC is available from http://gateway.ebioscience.amc.nl/ and can submit to Dutch gird and local clusters.

The SOMNO.NETZ community were interested in starting their analysis workflows from the XNAT interface (Extensible Neuroimaging Archive Toolkit, http://www.xnat.org/). XNAT is one of the most popular tools being widely adopted by neuroscientists for data management. It provides a storage server specialized for neuro-imaging research, and it allows invocation of external pipelines to further process these images. Although it was primarily for medical image data, it is currently also being exploited for data analysis, with possible scaling up capabilities. SOMNO.NETZ created a dedicated installation of the generic WS-PGRADE system and configured it for submission of jobs to their private Openstack cloud resources.

Most of the workflows can also be started from the SHIWA Portal using the test data available on the SHIWA Repository. This usage scenario is meant for teaching, dissemination/publication, and workflow sharing with colleagues outside the AMC. At the moment, we do not foresee biomedical users from the AMC executing the workflows directly from the SHIWA Portal for their own data due to privacy and usability considerations.

12 Application Porting Experience

The introduction of meta-workflows and the increased number of workflow engines used during Year 2 highlighted different usage patterns by the different communities and, thus, different porting experiences.

The **Computational Chemistry** (MoSGrid) community found that the porting of standard WS-PGRADE workflows to the SHIWA Repository has taken place without problems, and that the transfer of these workflows to the SHIWA Simulation Platform went quite fine. The annotation of additional data was easily processed on the SHIWA Repository. This community experienced that, with meta-workflows and meta-meta-workflows, the process was more complicated due to some errors, but the SHIWA team has resolved this in less than one day normally. The MoSGrid community also had the ambition of porting Galaxy and UNICORE workflows to the platform, however these were less successful. The porting of Galaxy workflows to WS-PGRADE was very difficult, and performed in the end manually by N. Weingarten. Since all MoSGrid workflows are based on the UNICORE middleware, support for this system was necessary. However it was not available in the SHIWA platform at the start of the project. The SHIWA support team worked on this very hard, and the problem was solved only very recently. Therefore, it was not possible to explore this new feature of the SHIWA platform before the preparation of this deliverable. Concerning the communication between the portals and the SHIWA Repository, it would be more practical if the SHIWA Repository password would be saved at the portal and automatically generated during import/export of workflows (single sign-on).

The **Astrophysics** and **Heliophysics** community focused on porting TAVERNA workflows and their combinations into meta-workflows in Year 2. These communities found that the porting of non-native workflows to the repository highlighted some aspects of the interface that were cumbersome and repetitive (thus more prone to errors). To overcome these difficulties, the communities have reported their suggestions and concerns to the University of Westminster to increase the efficiency of the interface and to reduce the redundancy of the information requested. This close interaction between the communities and the technology provider has driven the development of the SHIWA Submission Service releases of Year 2, which have ameliorated many of the highlighted issues. The Astrophysics community has also experienced problems related to the Virtual Organization management. Information systems were not well configured or unreliable, which was difficult to troubleshoot from the SHIWA Portal system administrator's perspective. These issues were sorted out by the Astrophysics community together with WP3. The Heliophysics community has also highlighted difficulties in describing the relationship between the meta-workflow and sub-workflows in the repository. The concept of meta-workflow is powerful and adopted by the communities, but it is somehow not seamlessly supported by the infrastructure.

The **Life Sciences** community has successfully ported new applications and improved the implementation of existing ones. In particular it has introduced the CERN Virtual Machine File System (CVMFS) as a software repository to enable running applications on a variety of infrastructures (Grid, Cloud, etc.) much more easily. WS-PGRADE was useful in adopting CVMFS, as it allows us to create global configurations at DCI-BRIDGE level that affect all workflows and are needed for usage of CVMFS. The community experienced that some features are still missing in the SHIWA platform for WS-PGRADE workflows, as detailed below.

Exporting a WS-PGRADE workflow to the SHIWA Repository has become straightforward, as a simple operation from the portal interface. As a second step, the documentation is input manually into the repository, both about the application executed by the abstract workflow,

and about the technical details about the workflow implementation(s). There is much interest in reusing the documentation by grouping workflows that run the same application in the repository. However, it has been difficult to do this in practice. Specifically, the communities have improved the implementation of some of the workflows that were already in the repository, and which were already documented. When the new workflows are uploaded into the SHIWA Repository via the WS-PGRADE export operation, a new abstract workflow and a new concrete workflow (or implementation) is created automatically. This requires either a manual step by SHIWA Repository system administrator to merge the two implementations (the old and the new one) under the same abstract workflow, or duplicating the documentation.

The Life Sciences community also experienced inconvenience caused by missing functionality to describe workflows in WS-PGRADE. An example is workflow loop, which is needed for one of the neuroimaging analysis applications: the DTI Population Registration workflow consists of an iterative process to refine results. In MOTEUR this was implemented as a controlled loop; however, in WS-PGRADE it had to be implemented as a fixed sequence of six repeated steps. This community also faced some problems in combining existing WS-PGRADE workflows into bigger meta-workflows. On one hand, using the WS-PGRADE native solution is not possible for workflows that use remote files as inputs: it turns out that currently hierarchical workflows are possible only when the data is uploaded/downloaded via the portal. Unfortunately the biomedical applications considered in ER-flow could not use this simpler approach because imaging and DNA sequencing data are very large. On the other hand, the SHIWA platform currently does not allow for the so called “black-box” execution of WS-PGRADE sub-workflows which uses the SHIWA CGI solution. As a consequence, it is not possible to make meta-workflows using the CGI technology either. Therefore, reusing workflows to compose more complex ones would require modifications of the data access and therefore hamper actual reuse.

And finally, the Life Science community experienced difficulties to reconfigure the data storage resources to be used by the workflows. The VLEMED storage elements used by the Life Sciences community have been very unreliable for a period of over a month, which greatly hindered the porting process and the customized gateways operation. The positive point about WS-PGRADE is that whenever a workflow execution would fail due to such instabilities, the workflows could be resumed without the need of further configuration. The negative side is that changing the preferred storage element at workflow level is not very easy, especially for workflows that have been put into production from customized interfaces of science gateways. This would entail changing the configuration of every output in every workflow, redeploying them into WS-PGRADE local repository, and reconfiguring the science gateway databases to take the new workflows.

The Life Science community also supported the SOMNO.NETZ project in a study to evaluate the applicability of WS-PGRADE to implement a new cloud-based portal. This experience delivered a useful evaluation of the WS-PGRADE and gUSE technologies, which were reported to WP3 and the WS-PGRADE developers. The project participants reported the following advantages of using WS-PGRADE/gUSE:

- Provides out-of-the-box cloud management.
- Can limit the maximum number of VM instances for improved resource allocation.
- Can reuse existing VMs for data processing, this reducing overhead.
- Workflows may be useful for large applications (but not for small sample applications)

Nevertheless, some problems were encountered during this experiment, which would in summary translate to the fact that the cloud implementation of WS-PGRADE is not mature enough yet, as it is not widely tested in different scenarios. The following points of further improvement have been reported:

- Data management is not optimized for cloud
  - always transfers data back to server
  - may process two jobs of one app with two VMs
- Rest API is not working out of the box for cloud
- Cloud workflows sometimes refused to start or data transfer sometimes did not work. In such cases, a restart of gUSE was necessary.
- Only manageable from graphical portal interface (no command line interface)
- Cannot deal properly with problems caused by the cloud middleware
- Insufficient error reporting (and not enough information in logs)

In summary, the porting and execution of native workflows has highlighted no major issues, since this technology has been used successfully in a large number of applications already in various scientific domains. Also, the WS-PGRADE technology has been improved during the second project year, as a result of efforts from both ER-flow and SCI-BUS projects. Nevertheless, the communities still reported various suggestions for further improvement and missing features. We highlight a few below:

- Unified password management (single sign-on)
- Visualization of managed VMs and job queue
- User-friendly web interface
  - usage of checkboxes should be improved.
  - save and upload workflow buttons should be made more accessible.
  - gUSE needs restart after changing settings
- Cloud debugging tools
  - Allow not shutting down VM to give the possibility of investigating bugs
  - Fully functional Remote API / command line interface

Moreover, all the communities have tested the use of the new SHIWA Submission System through the SHIWA Portal operated by the University of Westminster. Attempts to connect community portals to this submission service highlighted issues that prevent successful submission from portals behind a proxy. It is also worth mentioning that the execution environment is very complex and may expose unforeseen problems such as proxy and firewall restrictions that may block set of web services that use specific port numbers.

And finally, the experiences of three communities regarding the usage of workflow management systems as the back-end of science gateways has been presented in a paper and presentation at the IEEE e-science conference in Oct 20-24, 2014:

> Silvia Olabarriaga, Gabriele Pierantoni, Giuliano Taffoni, Claudio Vuerli, Giuliano Castelli, Eva Sciacca, Mohammad Mahdi Jaghoori, Vladimir Korkhov, Ugo Becciani, E Carley and Bob Bentley. Scientific Workflow Management - for whom? In: Proceedings of the IEEE e-Science Conference, Guaruja, BR. October 2014

13 Conclusions

In this deliverable we have described the applications ported and developed in the second year of the project. They represent the efforts of four communities to port relevant scientific applications to run on various distributed infrastructures using workflow management technology. These applications can be now executed from large variety of environments, including the SHIWA Simulation Platform and customized interfaces of community science gateways.

The large heterogeneity among the communities regarding workflow development, and their needs already observed in Year 1, has resulted in different approaches in the second year. Depending on the most common requirements of its users, the different communities developed approaches that range from hiding workflows behind customized portlets for a specific scientific research case, to developing multi-layered workflow architectures to foster re-usability, and to taking full advantage of workflow interoperability allowing their users to focus their efforts on their languages of choice.

Concerning **technological relevance**, we observed that each community has followed the assessment of the current state-of-the art of workflows performed in Year 1, and accordingly each one has implemented its own strategy for the creation, adaptation and porting activities. This heterogeneous but optimized strategy in each case has led to the successful porting of a large number of applications to the SHIWA Simulation Platform. Some communities focused on workflows expressed in the native environment of the SHIWA Portal (WS-PGRADE), while others have devoted their efforts to porting workflows developed for and from other platforms – primarily TAVERNA and Galaxy so far. Meta-workflows have been explored for experiments that would require much logistics without the SHIWA platform. Although it is still necessary to continue to improve and continue to support the SHIWA workflow interoperability technology, it is clear from the large amount and variety of cases presented in this deliverable that this approach is useful and promising.

Concerning **scientific relevance**, we observed that each community has selected a set of workflows that were meaningful to the scientific community and has demonstrated so by the implementation of a large number of Science Cases. As expected, the variety of science cases is large, considering the large number of research domains directly and indirectly involved in this project.

At the end of this two-year long effort, it has been interesting to observe the variety of usage scenarios that each community has developed to tackle their scientific challenges. The platform has offered a system flexible and configurable enough to support scientist in the access of very diverse resources ranging from web services to grids and, even more importantly, in defining custom tailored strategies in the design, implementation and execution of workflows across different languages and management systems. This legacy of solutions is almost as important as the workflows themselves.

Annexes

Details about the workflows implemented in the second year are presented in separate documents, one for each user community. Normally the annexes are included in the same document, however in this case this approach would make the document too long. We therefore chose to separate the annexes into individual files as follows:

- Annex A: Astrophysics – Description of Applications Ported to the SSP in Year 2
  File name: D5.5.Annex.A.Astro.pdf

- Annex B: Computational Chemistry – Description of Applications Ported to the SSP in Year 2
  File name: D5.5.Annex.B.CompChem.pdf

- Annex C: Heliophysics – Description of Applications Ported to the SSP in Year 2
  File name: D5.5.Annex.C.Helio.pdf

- Annex D: Life Sciences – Description of Applications Ported to the SSP in Year 2
  File name: D5.5.Annex.D.LifeSciences.pdf