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*5.2*
Description of applications ported to the SSP (year 1)

Due date of deliverable: 28/08/2013
Actual submission date: --/--/2013
Start date of project: 01/09/2012
Duration: 24 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

2 List of Figures

3 List of Tables

4 Status and Change History

Status:	**Name:**	**Date:**	**Signature:**
Draft:	Gabriele Pierantoni	30.08.2013	n.n. electronically
Reviewed:	Gabor Terstyanszky Gergely Sipos Karolis Eigelis	25.09.2013	n.n. electronically
Approved:			n.n. electronically

Table 1, Deliverable Status

Version	**Date**	**Pages**	**Author**	**Modification**
0.1	13/05/13	All	GP	Created Skeleton
0.2	23/05/13	All	GP	Changes to address suggestions. Added specific sub-section for each application and Usage section for each Community
0.3	06/08/13	All	GP	Added Astrophysics contribution. Draft for review by the project members.
0.4	08/09/13	All	GP	Completed Helio contribution and added Computational Chemistry and BioMed contributions
0.5	19/09/13	All	GP	Draft sent out for corrections.
1.0	24/09/13	All	GP	Draft of final version
2.0	28/09/2013	All	GP+SDO	Final version

Table 2, Deliverable Change History

5 Glossary

SSP	SHIWA Simulation Platform
WP	Work package

Table 3, Glossary

6 Abstract

This deliverable describes (both from a technical point and scientific point of view) the applications ported to the SHIWA platform during the 1st year of activities of Work Package 5. This constitutes the main technical activity of the ER-FLOW project. The deliverable contains background about the scientific domains addressed by the applications; the relevance of the selected applications in their respective fields; 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.

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 Application Support that deals with the creation and porting of applications of four different communities to the SHIWA Simulation Platform. This deliverable (D5.2) is issued at the end of first project year and aims at describing the porting process in this period.

The porting process of 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 if the application is still active or is out of date, in case the application needs updates or changes, and to update them prior to the porting process.
- The porting process itself, which entails development or adaption of workflow(s) that implement the application, their publication in the SHIWA repository, and their execution from one of the various SHIWA 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 did we port these applications**: A concise report of the process, the experience and issues related to the porting process. In fact deliverable D5.1 extensively document this experience, so here we only emphasize aspects that are relevant for each of the specific applications.

This deliverable has been structured with the aim of facilitating direct access to information at different levels of details in the document. The level of details in this deliverable are:

- Document-wide: information that pertains to the entire document:
  - ***Section 7, Introduction***: context information of the Deliverable in WP5 of ER-FLOW.
  - ***Section 12, Conclusions:*** presents general conclusions on the porting applications under WP5 or ER-FLOW in the first year of activity.
- Community-wide: information that applies to an entire community is available in the first sections for each community:

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

- ***Sections x.1, Ported Applications:*** This section describes what is common to all the ported applications within a given community.
- ***Sections x.2, Applications Usage:*** This section describes common features on how applications are used in the community. Note that in most cases the customized science gateways under construction by the four communities can provide a more friendly interface from which the scientists can run these applications. These gateways are built in the scope of the SCI-BUS project; therefore, an MoU was signed between ER-flow and SCI-BUS to stimulate the usage of SHIWA technology in these gateways and to exploit the applications ported as workflows more easily from these gateways.
- ***Sections x.3, Technical Background:*** This section describes common characteristics of the technical background of the applications of a given community, such as middleware, workflow languages and systems, etc.
- ***Sections x.5, Porting Experience:*** This section summarizes the porting experience of the community

- Application-wide: each application is described in detail in a separate section:
  - ***x.4.y.1, Nature and Relevance:*** This section describes features and relevance of the ported application in the scientific domain.
  - ***x.4.y.2, Usage:*** This section describes the common usage patterns of the ported application, and the environments from which it is executed.
  - ***x.4.y.3, Software Details:*** This section covers details of the software used by the application.
  - ***x.4.y.4, Workflow Details:*** This section covers details of the workflow.
  - ***x.4.y.5, Further Technical Details:*** This section 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 description templates for all the applications are available on the ER-FLOW web site (http://www.erflow.eu/applications).

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 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 behaviour. The studied objects include galaxies, stars, planets, extra-solar planets, the interstellar medium and the cosmic microwave background. Their emissions are examined across all parts of the electromagnetic spectrum, and the examined properties include luminosity, density, temperature, and chemical composition. Because astrophysics is a very broad subject, *astrophysicists* typically apply many disciplines of physics, including mechanics, electromagnetism, statistical mechanics, thermodynamics, quantum mechanics, relativity, nuclear and particle physics. In practice, modern astronomical research involves a substantial amount of physics.
Generally, either the term “astronomy” or “astrophysics” may be used to refer to this subject. Based on strict dictionary definitions, “astronomy” refers to “the study of objects and matter outside the Earth's atmosphere and of their physical and chemical properties” whereas “astrophysics” refers to the branch of astronomy dealing with “the behaviour, physical properties, and dynamic processes of celestial objects and phenomena”. In some cases, “astronomy” may be used to describe the qualitative study of the subject; instead, “astrophysics” is used to describe the physics-oriented version of the subject.
Cosmology is the study of the origins and eventual fate of the universe. Physical cosmology is the scholarly and scientific study of the origin, evolution, structure, dynamics, and ultimate fate of the universe, as well as the natural laws that keep it in order. Modern cosmology is dominated by the Big Bang theory, which attempts to bring together observational astronomy and particle physics. Cosmology is also connected to astronomy, but while the former concerns the Universe as a whole, the latter deals with individual celestial objects.
Stellar evolution is the process by which a star undergoes a sequence of radical changes during its lifetime. Depending on the mass of the star, this lifetime ranges from only a few million years for the most massive to billions of years for the least massive. Stellar evolution is not studied by observing the life of a single star, as most stellar changes occur too slowly to be detected, even over many centuries. Instead, astrophysicists come to understand how stars evolve by observing numerous stars at various points in their lifetime, and by simulating stellar structure using computer models.
Astroparticle physics is a branch of particle physics that studies elementary particles of astronomical origin and their relation to astrophysics and cosmology. It is a relatively new field of research emerging at the intersection of particle physics, astronomy, astrophysics, detector physics, relativity, solid state physics, and cosmology. Partly motivated by the historic discovery of neutrino oscillations, the field has undergone remarkable development, both theoretically and experimentally, over the last decade.

7.1 Ported Applications

The applications the Astro community brought on the SSP during the first year of the ER-flow project deals with Astronomy, Astrophysics, Cosmology, Stellar evolution and Astroparticle physics. Such applications have been selected partly because they were recognized as good representatives (demonstrator applications) in their own research field and partly because their porting on specific DCIs were already investigated and attempted in the past; this preliminary porting activity generated DCI-compatible, although not complete, versions of such applications, so the ER-flow project provided the right context allowing to exploit the experience gained through this past activity and to carry out the complete porting activity of selected applications. The most part of research groups that have provided the applications chosen for the first year are scientists with limited or no knowledge at all of the technology used. A large effort therefore was spent by people leading the astronomical participation in ER-flow to provide the technological support to successfully complete the porting activity.
The astrophysical applications ported on SSP during the PY1 are summarized in Table 4.

Workflow	Description
COMCAPT	Captures the trajectories of comets from the interstellar space
FRANEC/ BaSTI	FRANEC generates stellar evolutionary models and saves them in the BaSTI database
LasMoG	Produces customised visualisation for analysis of modified General Relativity (GR) simulations
MESTREAM	Produces simulations for the study of meteor showers situated in the orbital phase space of the orbit of asteroid 2003 EH1.
Planck	Implements the simulation of the LFI instrument of the ESA Planck space mission
VisIVO	The application produces a 3D movie representing the evolution of a cosmological N-body simulation into the defined sub-region

Table 4, Astrophysics Workflows Overview

7.2 Applications Usage

The workflows generated for the ported applications are executed in the same way. There are four possible access modes under development in the project:

- Through the **workflow developers interface** of the WFMS, in this case a generic installation of the WS-PGRADE portal available for members of the Astrophysics community. This access mode has been adopted during the porting and test phases for all applications.
- Through the **SHIWA Simulation Platform (SSP)**. This access mode as well is usually adopted during the porting and testing phase.
- Through the specialized science gateways that are made available by the SCI-BUS project to the Astrophysics community, allowing end users to easily find, retrieve and run applications and workflows they need to carry out their scientific goals.
- Through a **Science Gateway of the STARnet federation**, which is a network of astronomical science gateways where each of them is built on top of a generic common science gateway and provides a number of functionalities and services related to one or more specific applications or workflows. The basic common gateway makes available a set of basic features and services, including a set of those available in the central SSP, which are used by all applications and workflows. Science Gateways of STARnet are then used to look for and execute all astronomical applications and workflows.

7.3 Technical Background

Some of the astronomical applications selected for SSP during the first year of ER-flow did not have an existing workflow associated until now; for them, a workflow was developed using the gUSE/WS-PGRADE system; for other applications a WS-PGRADE native workflow was developed in the past. Therefore, the challenge of combining workflows coming from different workflow systems was not faced by the astronomical community during the first year. This situation however will change during the second year, with workflows brought in ER-flow coming from different workflow systems, and then the SSP will have to be used to its full extent to overcome these new challenges.

Moreover, some of the six monolithic workflows proposed for the first year are now in the process to be split in different sub-workflows providing basic functionalities useful for other applications and for their workflows; they are therefore building blocks that can be re-used for different scientific applications.

Some of the applications will be installed locally at the various sites that contribute computational and storage resources. Applications installed locally typically require some other software packages and depend from them. In these cases, applications can be run whatever is the science gateway of the end user but their execution takes place remotely.

Several VOs now support astronomical applications in ER-flow. Some of these VOs are specifically dedicated to a given project or application; others have a more general purpose and are typically related to a given community and to their applications.

7.4 Applications Description

7.4.1 COMCAPT

7.4.1.1 Nature and Relevance

The COMCAPT application deals with Astrophysics. The trajectories of interstellar comets passing the Solar System are gravitationally influenced by the Galactic tide. A combination of this influence and gravity of the Sun can change the trajectories in the way that the comets become bound to the Solar System, i.e. they become a part of the comet Oort cloud. For the current position of the Sun in the Galaxy and considering its relatively high peculiar velocity, the intervals of the comet orbital phase space, where the “capture” happens, occur to be extremely narrow. In addition, a preliminary analysis revealed the non-linear nature of the problem. So, the appropriate “capture window” can appear for an unexpected combination of comet orbital parameters (one cannot simply look for a mathematical local minimum).

7.4.1.2 Software Details

COMCAPT calculates the critical parameters of the capture for a huge number of interstellar-comet trajectories (of order of magnitude equal to 10^{16}) and evaluates if the condition of the capture is satisfied for the given combination of 4-D orbital characteristics or not.
COMCAPT is expected to be re-run for various combinations of two input values:

- distance of the Sun from the Galactic centre and
- Magnitude of the peculiar velocity of the Sun with respect to the LSR (Local Standard of Rest).

From the computational point of view, the application is a parametric study. Using the input data and specific astronomical software (created by users), it calculates some critical parameters and, based on these parameters, evaluates if the expected phenomenon (capture of interstellar comets into the comet Oort cloud) happens for a given combination of input data.

7.4.1.3 Workflow Details

The COMCAPT workflow can be easily modified to adapt it for another parametric type application. The user (astronomer) writes the source code doing calculations in a studied scientific problem and creates the appropriate input data.
After having split data in N parts corresponding to the N available CPUs, the executable code and data can be brought on the UI (User Interface) and the workflow can be used to perform the computations required by the application. The output data from the extensive computation can be further processed using a common personal computer to create tables, figures, movies, etc. for graphical and tabular representations of the results of the study.

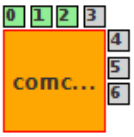

Figure 1, The COMCAPT Workflow Graph (WS-PGRADE)

7.4.1.4 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=4975
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-COMCAPT.pdf
User documentation	http://www.astro.sk/~mjakubik/WORKFLOWS/COMCAPT/
Software documentation	http://www.astro.sk/~mjakubik/WORKFLOWS/COMCAPT/
Contact details	Lubos Neslusane-mail: ne@ta3.sk Marian Jakubik e-mail: mjakubik@ta3.sk

Table 5, COMCAPT application technical details

7.4.2 FRANEC/BaSTI simulations

7.4.2.1 Nature and Relevance

The FRANEC/BaSTI application deals with stellar evolution and Astrophysics.
FRANEC is a state-of-art; numerical code for stellar astrophysics, this code is perfectly suited for computing the evolution of a star on the basis of a number of different physical inputs and parameters.
The BaSTI (Bag of Stellar Tracks and Isochrones) database is a theoretical astrophysical catalogue that collects fundamental data sets involving stars formation and evolution.
The BaSTI relational database is a suite of stellar evolution tracks, isochrones, luminosity functions and complementary codes to study the properties of resolved and unresolved stellar populations with an arbitrary SFH (Star Formation History).

7.4.2.2 Software Details

Parameters are listed in one input file. A single run of FRANEC produces one synthetic model (SM). To produce an isochrone, for a given chemical composition, through a FIR (Full Isochrone Run), it is necessary to execute a large number of SMRs (SM runs) varying the initial mass of the stellar models. Once these evolutionary tracks and isochrones (as well as additional data describing the simulated stellar structures) are computed, they can be distributed in datasets over different sites.
The simulations of stellar models produce simulation output files with a set of associated metadata. Such metadata are linked to all parameters concerning the numerical evolutionary code. In this way it is possible to store and easily search and retrieve the obtained data by many set of stellar simulations, and also get access to a huge amount of homogeneous data such as tracks and isochrones computed by using FRANEC.
All stellar model simulations and their characterizing parameters, the produced output files and their metadata and the relationships (links) between them are stored and maintained in BaSTI.
BaSTI allows in this way to archive and publish the data of many stellar evolution simulations; it also offers to the scientific community the possibility of reusing a large number of stellar model computations.
According to the planned evolution, the BaSTI database will be automatically updated when new simulations are available. BaSTI, moreover, could be updated on request. When an astronomer requests some data that are not available in the database, a service will allow to submit a FIR to update the database with the requested data. The service is moderated, i.e. the content managers decide if the requested data should be computed or not according to their scientific relevance/interest.

7.4.2.3 Workflow Details

The BaSTI/FRANEC workflow has a typical modular architecture; it is easy to identify its modules that can be reused to build other workflows. Modules can be identified on the basis of the function(s) they provide: 1) retrieval of the simulation files obtained (output) through past simulation runs[2]; 2) retrieval of a synthetic model simulation[3]; 3) new simulation of a synthetic model[4]; 4) ingestion of a new synthetic model[5]; 5) post-processing analysis[6].

[2] The end user specifies a set of metadata. Metadata of each simulation output file are compared against metadata provided by end user. At the first match occurrence, the corresponding output file is returned.

[3] Given a simulation output file, its associated metadata are used to retrieve and return to the end user the synthetic model simulation that originated the output file thanks to the relationship established between such metadata and the input parameters file.

[4] A new simulation of a synthetic model is performed. Skilled end users have to choose the synthetic model to simulate and the set of values for initial parameters. The scientific relevance of the new simulation has to be verified/certified by the Content Manager of the BaSTI database. If not relevant, the new simulation is

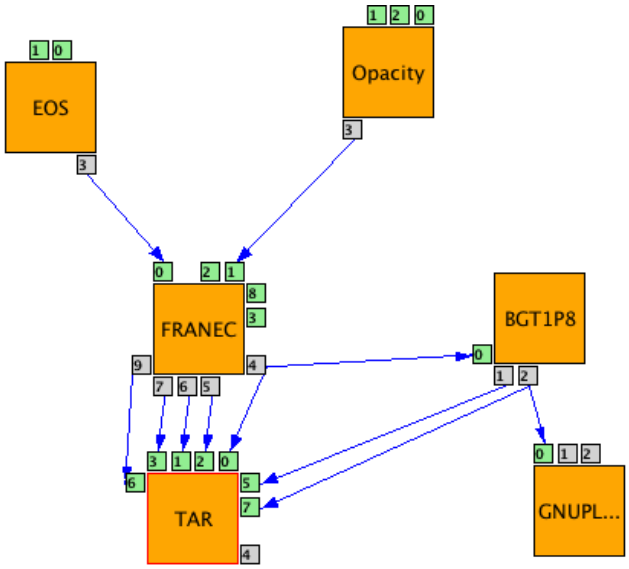

Figure 2, The FRANEC/BaSTI Workflow Graph (WS-PGRADE)

7.4.2.4 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=4979
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-BaSTI.pdf
User documentation	http://albione.oa-teramo.inaf.it/
Software documentation	http://albione.oa-teramo.inaf.it/
Contact details	Santi Cassisi e-mail: cassisi@oa-teramo.inaf.it Adriano Pietrinferni e-mail: pietrinferni@oa-teramo.inaf.it

Table 6, FRANEC/BaSTI application technical details

rejected. The resulting output file is stored in BaSTI together with its related metadata set up by the skilled end user.

[5] New synthetic models might be proposed by skilled end users to be inserted in BaSTI. The scientific relevance of the new synthetic model has to be verified/certified by the Content Manager of the BaSTI database. If not relevant, the new synthetic model is rejected.

[6] Final products resulting from synthetic model simulations and contained in the simulation output files might be used to perform specific post-processing analysis by running software packages provided by end users or third-party contributed. The necessary precondition is the deployment of such software packages on the adopted DCI.

7.4.3 LaSMoG

7.4.3.1 Nature and Relevance

The LaSMoG application deals with Astrophysics.
To understand acceleration of the universe, a weird component is introduced, called dark energy, in the framework of General Relativity (GR). GR needs to be tested on cosmic scales; previous studies show that cosmic acceleration could be realised by modifying GR on cosmological scales without introducing dark energy.
If GR gets modified, the structure formation will be very different from that in GR, although the expansion history remains the same as in the LCDM (λ-Cold Dark Matter) model of the universe (i.e. the standard model of big-bang cosmology). Observing the large scale structure of the universe could in principle provide new test of GR on cosmic scales. This kind of test cannot be done without the help of simulations as the structure formation process is highly non-linear. Large-scale simulations are performed for modified gravity models within the Large Simulation for Modified Gravity (LaSMoG) consortium.

7.4.3.2 Software Details

The application produces customised visualisation for analysis of modified GR simulations, more specifically inspecting datasets to discover anomalies by comparing appropriately with datasets coming from standard gravity models.

7.4.3.3 Workflow Details

Scientists are interested in comparing features within regions of interest from original time steps, then tracking such differences throughout a large sequence of snapshots. The users will submit the workflow configuring the input data files and parameters by an easy to use interface (portlet) via a science gateway.

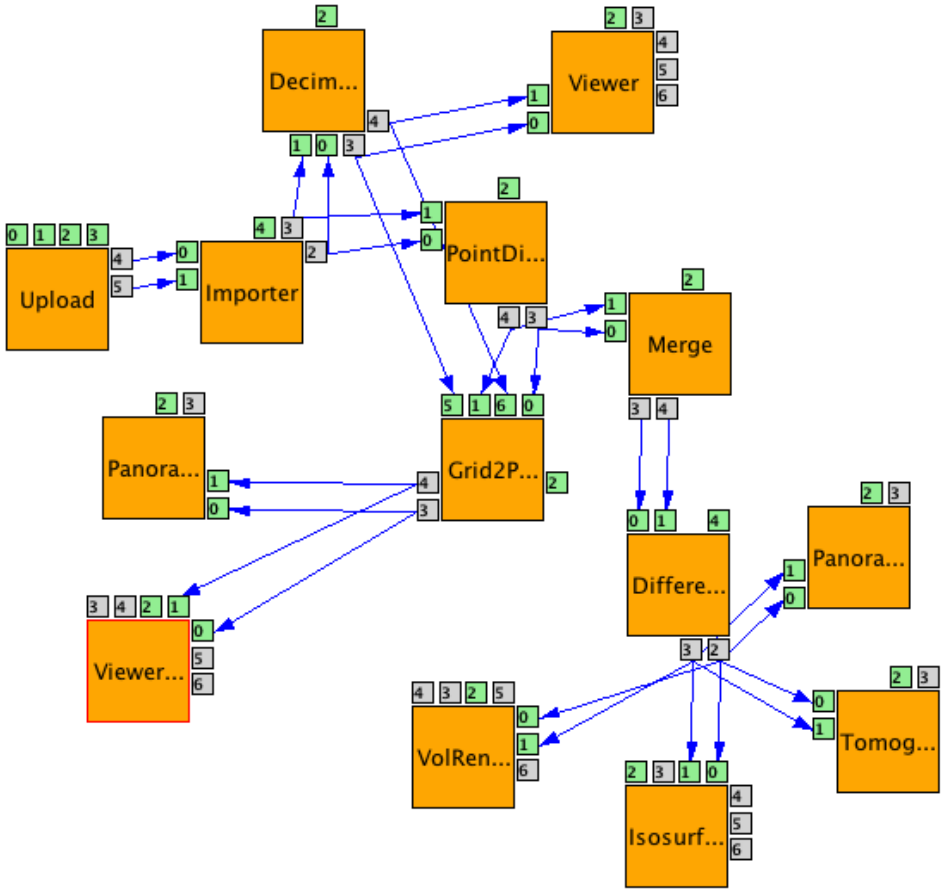

Figure 3, The LasMoG Workflow Graph (WS-PGRADE)

7.4.3.4 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=4980
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-LaSMoG.pdf
User documentation	http://www.itp.uzh.ch/~teyssier/Site/RAMSES.html
Software documentation	http://www.itp.uzh.ch/~teyssier/Site/RAMSES.html
Contact details	Mel Krokos e-mail: mel.krokos@port.ac.uk

Table 7, LasMoG application technical details

7.4.4 MESTREAM

7.4.4.1 Nature and Relevance

This application deals with Astrophysics.
The aim of the planned simulation is the study of meteor showers situated in the orbital phase space of the orbit of asteroid 2003 EH1. The real showers to compare the simulation results are mainly in 3 databases of the IAU Meteor Data Centre (photographic, video, and radio-meteor databases), which are gradually enlarged. The photographic database is managed by a team at our institute; this team releases a new version at “Meteoroids 2013” conference in Poznań, Poland. Another problem is represented by the outbursts of comet 29P/Schwassmann-Wachmann 1. The list of these events is known, but the actual reason is still under investigation. One possible reason are the collisions with the meteoroid particles released from the comet itself. In both cases, we intend to simulate the process of creation of the meteoroid stream and study its dynamical evolution. The executable codes and scripts to perform the computations are ready and were used in a similar application.

7.4.4.2 Software Details

After a (manual) emplacement of the input data and executable code from the user interface (UI) to the storage element (SE), these data and code are distributed, through a control script, to the individual computing elements (CEs) in the Grid environment. After the completion of this task, the output data are returned, by the script, back to the SE. The completeness of the output can be checked manually. When all is done, the user moves the output data from the SE to the UI (and, eventually, further to an archive medium).

7.4.4.3 Workflow Details

The workflow enables performing an extensive task (description of a model evolving in time), which can be divided into N independent sub-tasks.
The MESTREAM workflow can also be used to study the dynamical evolution of any population of small bodies in the inner Solar System (meteoroid streams of another several decades of parent bodies, asteroids in the main belt, Halley-type comets, Jupiter Trojans, objects in the Kuiper belt, scattered disk beyond Neptune, etc.). The user (astronomer) only specifies the input data (the position and velocity vectors) about the small bodies intended to be studied. These extensive data are required to be divided to N parts corresponding to N available CPUs, the executable code and data can be emplaced on the UI and the workflow can be used to perform the computations required by the application. The output data from the extensive computation are, then, finally processed using a common personal computer to create the tables, figures, movies, etc. which describe the result of the study.

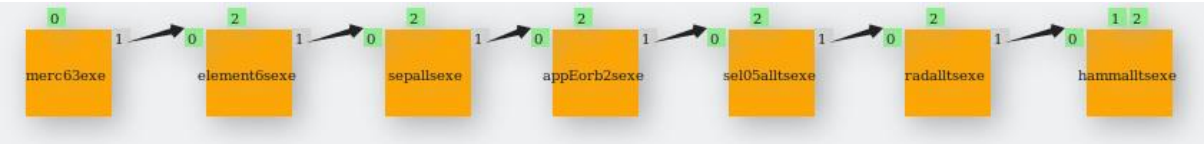

Figure 4, The MESTREAM Workflow Graph (WS-PGRADE)

7.4.4.4 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=4977
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-MESTREAM.pdf
User documentation	http://www.astro.sk/~mjakubik/WORKFLOWS/MESTREAM/
Software documentation	http://www.astro.sk/~mjakubik/WORKFLOWS/MESTREAM/
Contact details	Lubos Neslusane-mail: ne@ta3.sk Marian Jakubik e-mail: mjakubik@ta3.sk

Table 8, MESTREAM application technical details

7.4.5 Simulations of the Planck mission

7.4.5.1 Nature and Relevance

This application relates with Astronomy, Astrophysics and Cosmology.
The application implements the simulations of the Planck LFI mission. The workflow is designed on the requirements of applications that cannot be handled by a single computing farm, both in terms of computing power and data storage.

7.4.5.2 Software Details

The workflow consists of a very simple pipeline which is constituted 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 overlapping 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 and we can therefore 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.

7.4.5.3 Workflow Details

A simulation starts from a set of values for 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.

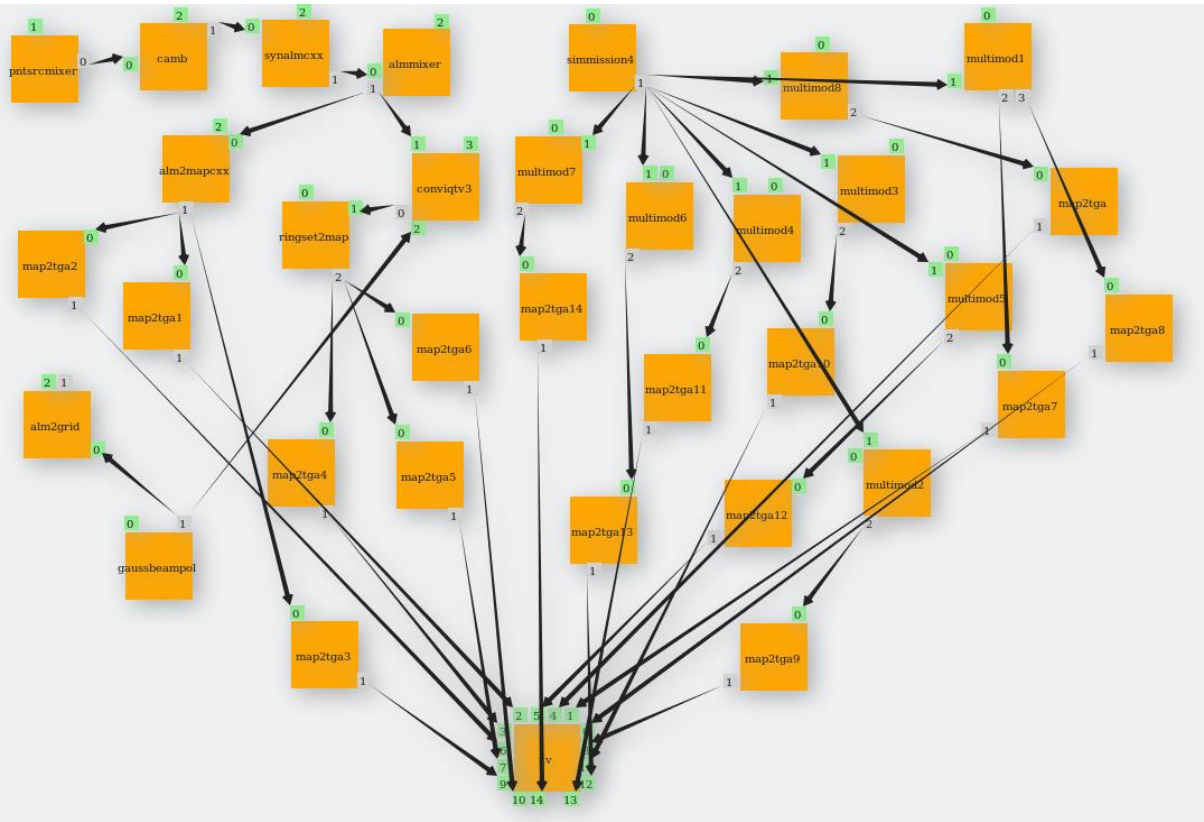

Figure 5, The PLANCK Simulations Workflow Graph (WS-PGRADE)

7.4.5.4 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=4978
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-Planck.pdf
User documentation	http://twiki.oats.inaf.it/twiki/pub/ADCIs/ErFlow/Planck-simulation-modules-usage.pdf
Software documentation	http://twiki.oats.inaf.it/twiki/pub/ADCIs/ErFlow/Planck-simulation-pipeline.pdf
Contact details	Giuliano Castelli e-mail: giuliano.castelli@oats.inaf.it Giuliano Taffoni e-mail: taffoni@oats.inaf.it Claudio Vuerli e-mail: vuerli@oats.inaf.it

Table 9, PLANCK Simulations application technical details

7.4.6 VisIVO

7.4.6.1 Nature and Relevance

The VisIVO application was initially designed and implemented as a tool to visualize astrophysical data. Thereafter it evolved becoming a general-purpose data visualization tool, suitable to be applied in different scientific domains and also for commercial and industrial purposes.

7.4.6.2 Software Details

Now VisIVO consists of a suite of software tools for creating customized views of 3D renderings from astrophysical data tables. These tools are founded on the VisIVO Desktop functionality (visivo.oact.inaf.it) and support the most popular Linux based platforms (e.g. www.ubuntu.com). Their defining characteristic is that no fixed limits are prescribed regarding the dimensionality of data tables input for processing, thus supporting very large scale datasets.
Assuming that datasets are uploaded, users are typically presented with tree-like structures (for easy data navigation) containing pointers to files, tables, volumes as well as visuals.
Files point to single, or possibly several (for distributed datasets), astrophysical data tables.
Tables are highly-efficient internal VisIVO Server data representations; they are typically produced from importing datasets uploaded by users using VisIVO Importer (see below).
Volumes are internal VisIVO Server data representations; they are produced either from direct importing of user datasets or by performing operations on already existing tables.
Visuals are collections of highly-customized, user-produced views of 3D renderings of volumes.
VisIVO Server consists of three core components: VisiVO Importer, VisiVO Filter and VisIVO Viewer respectively. VisIVO allows importing cosmological datasets and builds customized 3D visualization and movies from such datasets. A cosmological simulation produces a set of snapshots at different time steps with different time tags, not linearly distributed. The researcher is normally interested in sub-regions, voids or halos. The VisIVO application produces a 3D movie representing the evolution of a cosmological N-body simulation into the defined sub-region.

7.4.6.3 Workflow Details

The workflow will be accessed via VisIVO science gateway (http://visivo.oact.inaf.it:8080); the user will submit 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.

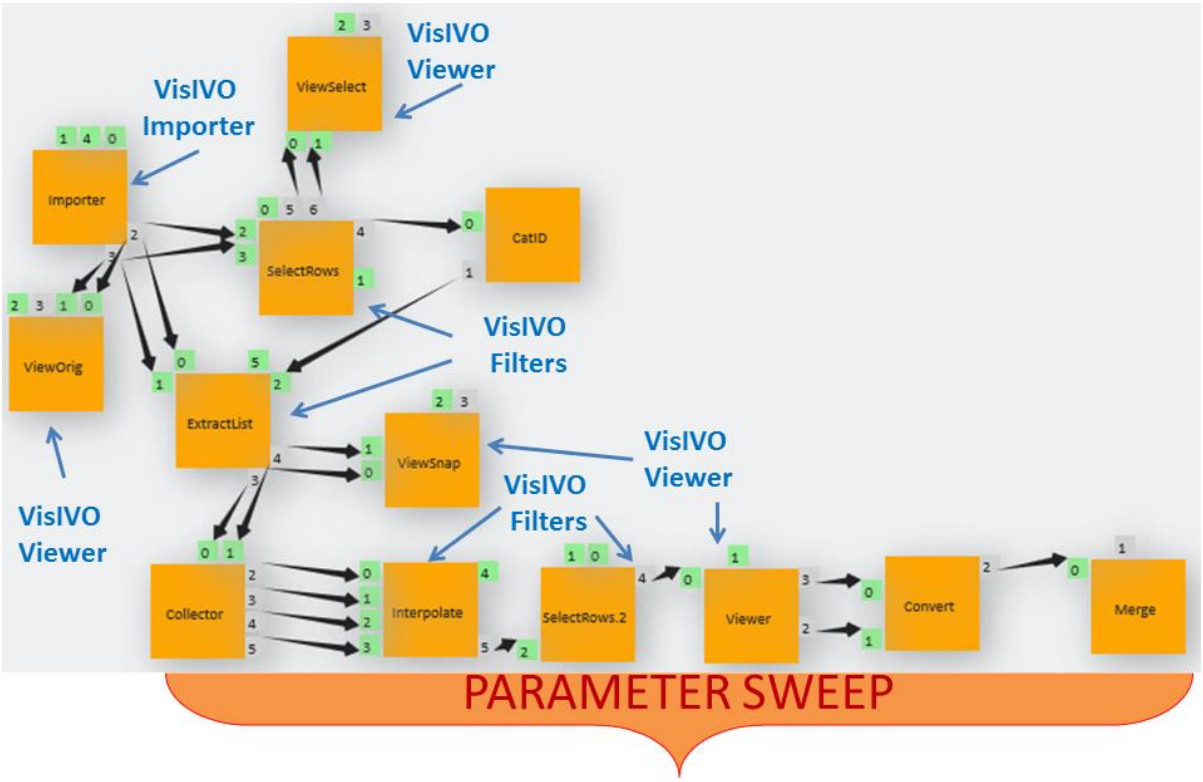

Figure 6, The VisIVO Workflow Graph (WS-PGRADE)

7.4.6.4 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=4961
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-VisIVO.pdf
User documentation	http://sourceforge.net/projects/visivoserver/
Software documentation	http://sourceforge.net/projects/visivoserver/
Contact details	Alessandro Costa e-mail: alessandro.costa@oact.inaf.it

Table 10, VisIVO Simulations application technical details

7.5 Porting Experience

The porting of astrophysical applications on the SSP and the creation/adaptation of workflows for this platform during the first year of ER-flow did not raise blocking issues. We met some minor problems that were reported to WP3 and were always solved in a short time.
In order to identify the astrophysical applications to be ported on SSP during the first year of ER-flow, face-to-face meetings were organized at the home institutions of the contributors who provided the applications for the first year of the project. These meetings allowed us to establish a fruitful tight collaboration between application contributors and DCI experts, which is a necessary precondition for successful application porting.

For what concerns the SSP, during the first year we made use of it mainly for development and testing purposes. All astronomical workflows ported on the SSP during the first year of the project are WS-PGRADE native. Therefore, for production purposes, some workflows have been submitted through our SCI-BUS gateway and run on the INAF cluster, and others workflows (FRANEC and COMCAPT) have been executed into the WestFocus infrastructure installed at the UoW (University of Westminster).

8 Computational Chemistry

The Molecular Simulation Grid (MoSGrid) is a German project which 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 on 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 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 [Ref: unicore] 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.

8.6 Ported Applications

For the first year, we have chosen applications which are representative for every domain of MoSGrid (Molecular Dynamics, Docking, and Quantum Chemistry). As generic approach, we have decided to use free simulation software. Within MoSGrid, we had already developed multiple workflows for quantum chemistry, but they were based on the expensive Gaussian09 software package. For further development of quantum chemistry workflows and application porting to SHIWA, we switched to the simulation code NWChem. For Molecular Dynamics, we use GROMACS and for Docking we used the code CADDSuite and AutoDockVina. All selected workflows are summarised in Table 11. The links to the ported workflows in the SHIWA repository are listed in Table 12.

Note that in MoSGrid we chose to document the workflows, both in this deliverable as in the SHIWA repository, using abstract visual representations, instead of 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.

COMPUTATIONAL CHEMISTRY				
GROMACS Molecular Dynamic application				
Energy Minimisation		WS-PGRADE	gLiTE/UniCore	MoSGrid VO
Equilibration		WS-PGRADE	gLiTE/UniCore	MoSGrid VO
Single TPR		WS-PGRADE	gLiTE/UniCore	MoSGrid VO
CADDSuite docking application				
Docking with ligand generation		WS-PGRADE	gLiTE/UniCore	MoSGrid VO
Docking without ligand generation		WS-PGRADE	gLiTE/UniCore	MoSGrid VO
AutoDockVinaFull		WS-PGRADE	gLiTE/UniCore	MoSGrid VO
NWChem Quantum Chemistry/Molecular Dynamics application				
Geometry optimisation = basic WF		WS-PGRADE	gLiTE/UniCore	MoSGrid VO
Opt+freq		WS-PGRADE	gLiTE/UniCore	MoSGrid VO
Freq WF		WS-PGRADE	gLiTE/UniCore	MoSGrid VO
TD-DFT WF		WS-PGRADE	gLiTE/UniCore	MoSGrid VO
Mulliken WF		WS-PGRADE	gLiTE/UniCore	MoSGrid VO
Solvation WF		WS-PGRADE	gLiTE/UniCore	MoSGrid VO
Spectroscopic WF		WS-PGRADE	gLiTE/UniCore	MoSGrid VO
Transition state WF		WS-PGRADE	gLiTE/UniCore	MoSGrid VO
TS analysis WF		WS-PGRADE	gLiTE/UniCore	MoSGrid VO

Table 11, Overview of ported MoSGrid workflows

GROMACS Molecular Dynamic application	
Energy Minimisation	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4355
Equilibration	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4354
Single TPR	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4857
AutodockVina docking application	
AutodockVinaFull	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4205
CADDSuite docking application	
Docking with ligand generation	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4803
Docking without ligand generation	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4802
NWChem Quantum Chemistry application	
Geometry optimisation = basic WF	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=3958
Opt+freq	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=3959
Freq WF	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4206
TD-DFT WF	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4751
Mulliken WF	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4753
Solvation WF	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4752
Spectroscopic analysis = Metaworkflow	postponed to second year
Transition state search	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4913
Transition state analysis = MetaWF	postponed to second year

Table 12, Links to ported MoSGrid workflows to the SHIWA repository.

8.7 Applications Usage

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

- Through the **SHIWA Simulation Platform (SSP)**. This modality is preferred by developers during the porting and testing phase.
- Through the **MoSGrid Developer Interface**. This modality is preferred by workflow developers who wish to combine workflows and develop new ones.
- Through the **MoSGrid User Interface** divided after domains. This modality is preferred by real 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.

8.8 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 manager. 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. The UNICORE grid middleware offers a complete stack of tools; a graphical user interface allows creating jobs and workflows and submitting 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. 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 in D5.3).

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.

8.9 Applications Description

8.9.1 GROMACS: Energy Minimisation

8.9.1.1Nature and Relevance

The workflow is intended for the new and inexperienced GROMACS user. The user has to provide a plain .pdb file containing the coordinates of a protein. After molecule selection, a force field and water model can be selected, followed by the number of energy minimization steps. Furthermore resource settings can be adjusted. Usually the default settings give reasonable results, producing a solvated and minimized simulation system within a couple of minutes (depending on cluster availability). Such a system can be used for long time simulation with GROMACS or may be processed further with other tools.

Force field based simulation tools like GROMACS work with so called topologies, describing the interactions within a protein with specific energy functions. Therefore only biomolecules that are described by the available force fields can be simulated. Pure proteins always work well, but if you want to simulate some fancy liquid crystals with strangely conjugated PI-systems you may run out of topology parameters.

8.9.1.2Usage

This workflow is used by logging in to the MoSGrid Science Gateway, then choosing the “Simulation” menu, “Molecular Dynamics”, “Gromacs”, and the “Energy Minimisation” workflow. After choosing input files and parameters the workflow is to be submitted. After the workflow is finished the results are available for download.

8.9.1.3Software Details

GROMACS is a Molecular Dynamics tool suite offering not only the main simulation application but also a broad variety of preparation and analysis tools. Typically a molecule e.g. protein is put into a simulation box, solvated and energy minimized. Then the productive simulation is carried out yielding a trajectory. This is then analysed for physical and geometric properties. The software is available under http://www.gromacs.org/, the usage is described under http://manual.gromacs.org/, and it can be cited via http://pubs.acs.org/doi/abs/10.1021/ct700301q

8.9.1.4Infrastructure Details

The MoSGrid Science Gateway is used to execute the workflow. It is based on Liferay, gUSE, and WS-PGRADE. It uses the underlying XtreemFS for storage and UNICORE to access connected clusters.

8.9.1.5Virtual Organization Details

The VO MoSGrid is used.

8.9.1.6Workflow Details

The input file format is PDB and the output file formats are XTC, TRR, PDB, GRO, and XVG. The input data sizes are between 1 and 50 MB, the output sizes are 0.05 to 10 GB, the memory usage is between 1 and 32 GB, and the disk usage is low. The processing time is between 5 min and several weeks. A graph can be found below.

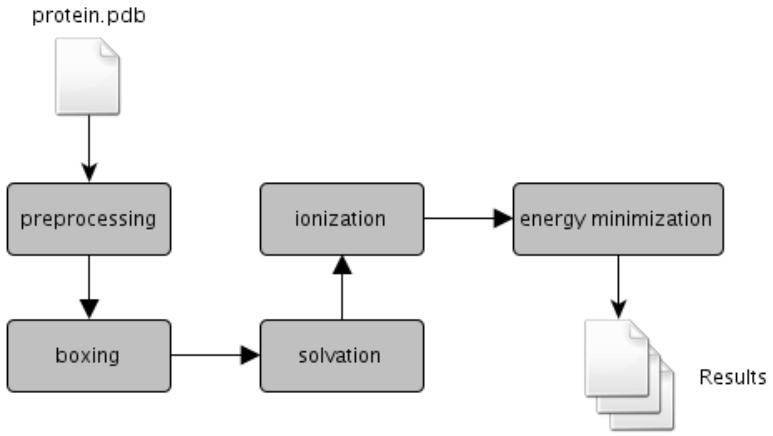

Figure 7, Gromacs Energy minimisation workflow (WS-PGRADE)

8.9.1.7 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=4355
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-GROMACS-EnergyMinimization.pdf
User documentation	http://manual.gromacs.org/
Contact information	Dr. Jens Krüger e-mail: krueger@bioinformatik.uni-tuebingen.de Richard Grunzke e-mail: richard.grunzke@tu-dresden.de

Table 13, Gromacs Energy minimization application technical details

8.9.2 GROMACS: Equilibration

8.9.2.1Nature and Relevance

The equilibration workflow corresponds to the energy minimization but offers the possibility to equilibrate the protein a little bit further. The minimization is carried out with position constraints on the protein, ensuring its integrity during this step. Afterwards a short constrained MD simulation is performed, allowing the solvent molecules to move freely. As last step a free MD simulation is carried out for a given simulation length. The protocol used for this workflow corresponds to the best practice used in the field.

8.9.2.2Usage

This workflow is used by logging in to the MoSGrid Science Gateway, then choosing the “Simulation” menu, “Molecular Dynamics”, “Gromacs”, and the “Equilibration” workflow. After choosing input files and parameters the workflow is to be submitted. After the workflow is finished the results are available for download.

8.9.2.3Software Details

GROMACS is a Molecular Dynamics tool suite offering not only the main simulation application but also a broad variety of preparation and analysis tools. Typically a molecule e.g. protein is put into a simulation box, solvated and energy minimized. Then the productive simulation is carried out yielding a trajectory. This is then analysed for physical and geometric properties. It is available under http://www.gromacs.org/; the usage is described under http://manual.gromacs.org/, and can be cited via http://pubs.acs.org/doi/abs/10.1021/ct700301q.

8.9.2.4Infrastructure Details

The MoSGrid Science Gateway is used to execute the workflow. It is based on Liferay, gUSE, and WS-PGRADE. It uses the underlying XtreemFS for storage and UNICORE to access connected clusters.

8.9.2.5Virtual Organization Details

The VO MoSGrid is used.

8.9.2.6Workflow Details

The input file format is PDB and the output file formats are XTC, TRR, PDB, GRO, and XVG. The input data sizes are between 1 and 50 MB, the output sizes are 0.05 to 10 GB, the memory usage is between 1 and 32 GB, and the disk usage is low. The processing time is between 5 min and several weeks. A graph can be found below.

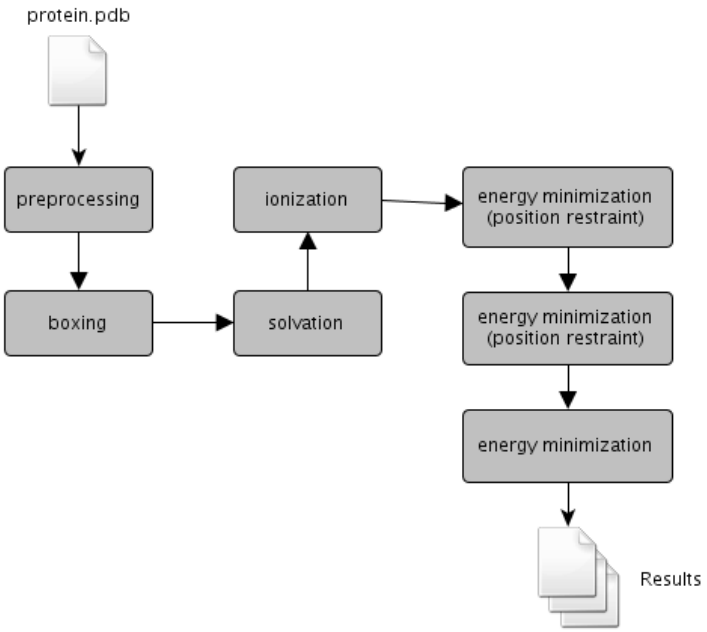

Figure 8, GROMACS Equilibration workflow (WS-PGRADE)

8.9.2.7 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=4354
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-GROMACS-Equilibration.pdf
User documentation	http://manual.gromacs.org/
Contact information	Dr. Jens Krüger e-mail: krueger@bioinformatik.uni-tuebingen.de Richard Grunzke e-mail: richard.grunzke@tu-dresden.de

Table 14, GROMACS Equilibration application technical details

8.9.3 GROMACS: Single TPR

8.9.3.1Nature and Relevance

This is a very basic workflow for experienced users, who mainly want to take advantage of the computational resources available through MoSGrid. It is assumed that the user is familiar with GROMACS and that a .tpr containing all information needed for a simulation has been prepared elsewhere. The user has to upload that file and specify the resource settings. In general all jobs are queued in the underlying grid system. The smaller the resource requirements are, the faster the job is started. It is strongly recommend to do a pilot run, setting the maximal runtime (corresponds to -maxh option of mdrun, i.e. the computational core of GROMACS) to a couple of minutes, in order to get an idea of the performance.

8.9.3.2Usage

This workflow is used by logging in to the MoSGrid Science Gateway, then choosing the “Simulation” menu, “Molecular Dynamics”, “Gromacs”, and the “Single TPR” workflow. After choosing input files and parameters the workflow is to be submitted. After the workflow is finished the results are available for download.

8.9.3.3Software Details

GROMACS is a Molecular Dynamics tool suite offering not only the main simulation application but also a broad variety of preparation and analysis tools. Typically a molecule e.g. protein is put into a simulation box, solvated and energy minimized. Then the productive simulation is carried out yielding a trajectory. This is then analysed for physical and geometric properties. It is available under http://www.gromacs.org/; the usage is described under http://manual.gromacs.org/, and can be cited via http://pubs.acs.org/doi/abs/10.1021/ct700301q

8.9.3.4Infrastructure Details

The MoSGrid Science Gateway is used to execute the workflow. It is based on Liferay, gUSE, and WS-PGRADE. It uses the underlying XtreemFS for storage and UNICORE to access connected clusters.

8.9.3.5Virtual Organization Details

The VO MoSGrid is used.

8.9.3.6Workflow Details

The input file format is TPR and the output file formats are XTC, TRR, PDB, GRO, and XVG. The input data sizes are between 1 and 50 MB, the output sizes are 0.05 to 10 GB, the memory usage is between 1 and 32 GB, and the disk usage is low. The processing time is between 5 min and several weeks. A graph can be found below.

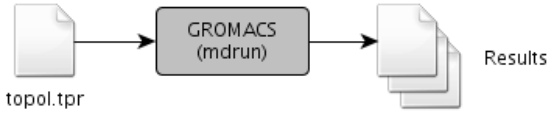

Figure 9, GROMACS Single TPR workflow (WS-PGRADE)

8.9.3.7 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=4857
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-GROMACS-SingleTPRWorkflow.pdf
User documentation	http://manual.gromacs.org/
Contact information	Dr. Jens Krüger e-mail: krueger@bioinformatik.uni-tuebingen.de Richard Grunzke e-mail: richard.grunzke@tu-dresden.de

Table 15, GROMACS Single TPR application technical details

8.9.4 Docking: AutodockVinaFull

8.9.4.1Nature and Relevance

This workflow performs a docking procedure of ligands into a protein. This version allows more advanced user input and is thus suitable for users who have performed docking before. The actual docking process includes splitting of the receptor-file into protein and reference ligand, docking with AutoDock Vina, and export of the dock results. AutoDock Vina creates a grid on the fly and thus no separate grid-building step is performed. For this workflow a protein structure file containing a protein in complex with a reference ligand and a file with the screening library (i.e., ligands that are to be docked into the protein) is required.

8.9.4.2Usage

This workflow is used by logging in to the MoSGrid Science Gateway, then choosing the “Simulation” menu, “Docking”, “AutoDock Vina”, and the “AutodockVinaFull” workflow. After choosing input files and parameters the workflow is to be submitted. After the workflow is finished the results are available for download.

8.9.4.3Software Details

AutoDock Vina is a docking tool. It is open-source and was designed and implemented by Dr. Oleg Trott. It is available under http://vina.scripps.edu/.

8.9.4.4Infrastructure Details

The MoSGrid Science Gateway is used to execute the workflow. It is based on Liferay, gUSE, and WS-PGRADE. It uses the underlying XtreemFS for storage and UNICORE to access connected clusters.

8.9.4.5Virtual Organization Details

The VO MoSGrid is used.

8.9.4.6Workflow Details

A graph can be found below.

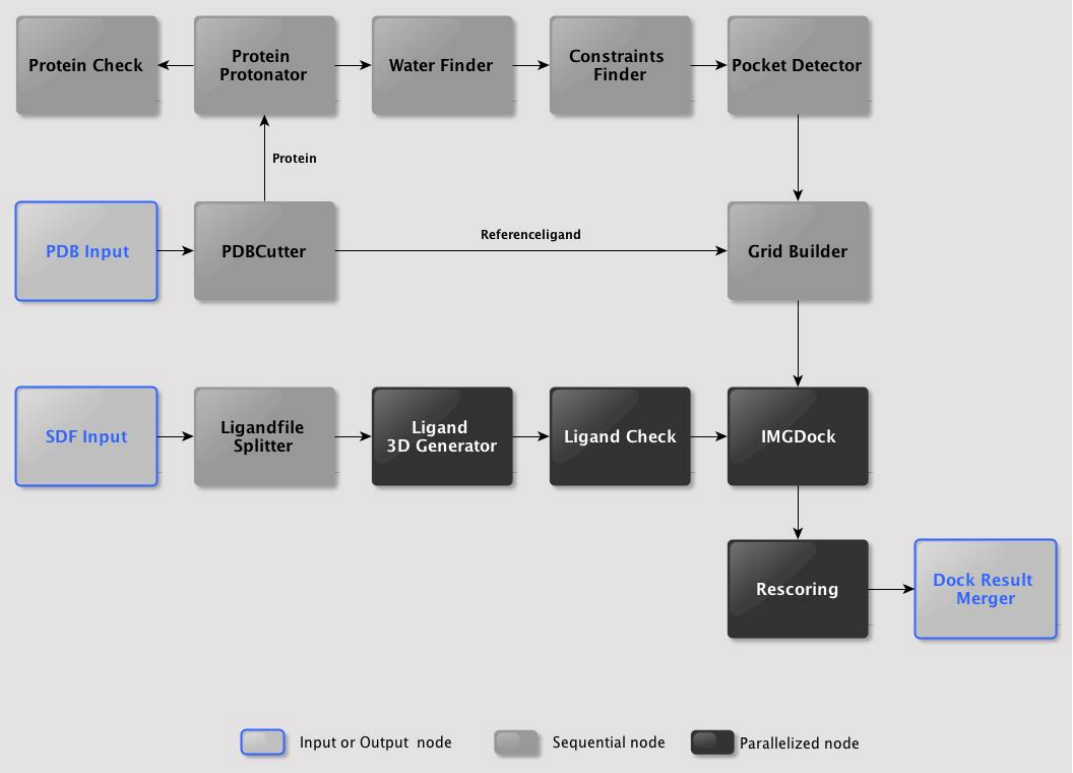

Figure 10, AutoDockVina Full workflow (WS-PGRADE)

8.9.4.7 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=4205
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-AutoDockVinaFull.pdf
Software documentation	http://vina.scripps.edu/
Contact information	Dr. Jens Krüger e-mail: krueger@bioinformatik.uni-tuebingen.de Richard Grunzke e-mail: richard.grunzke@tu-dresden.de

Table 16, AutoDockVina Full application technical details

8.9.5 CADDSuite: Docking with ligand generation

8.9.5.1Nature and Relevance

This workflow performs a docking procedure of ligands into a protein. This version is simplified so that only minimal user input is needed and thus is aimed at the novice user. The actual docking process includes splitting of the receptor-file into protein and reference ligand, grid building, docking and rescoring of the dock results. For this workflow a protein structure file containing a protein in complex with a reference ligand and a file with the screening library (ligands) is required.

8.9.5.2Usage

This workflow is used by logging in to the MoSGrid Science Gateway, then choosing the “Simulation” menu, “Docking”, “CADDSuite”, and the “Docking using CADDSuite” workflow. After choosing input files and parameters the workflow is to be submitted. After the workflow is finished the results are available for download.

8.9.5.3Software Details

The CADDSuite is a modular docking tool suite, offering various tools for the preparation, docking, scoring and analysis of protein ligand interactions. CADDSuite tools have been already used in Galaxy, gUSE, UNICORE and KNIME workflows. Using them up to multiple million ligands can be evaluated in short time.
It is available under http://www.ball-project.org/caddsuite, the usage is described under http://www.ballview.org/Support/caddsuite-tutorial-1, and it can be cited via http://link.springer.com/article/10.1186%2F1758-2946-4-S1-O2

8.9.5.4Infrastructure Details

The MoSGrid Science Gateway is used to execute the workflow. It is based on Liferay, gUSE, and WS-PGRADE. It uses the underlying XtreemFS for storage and UNICORE to access connected clusters.

8.9.5.5Virtual Organization Details

The VO MoSGrid is used.

8.9.5.6Workflow Details

The input file formats are PDB and SDF and the output file formats are PDB and SDF. The input data sizes are between 1 and 10 MB, the output sizes are about 100 MB, the memory usage is between 8 and 16 GB, and the disk usage is low. The processing time is about 10 seconds per ligand. A graph can be found below.

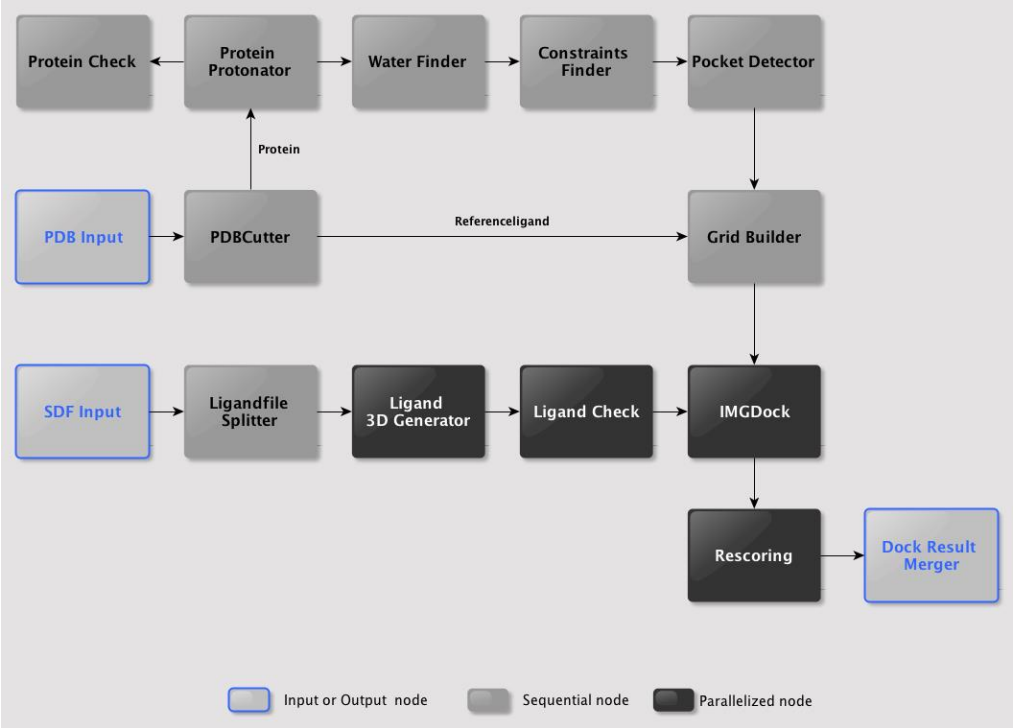

Figure 11, CADDSuite Docking with Ligand generation workflow (WS-PGRADE)

8.9.5.7 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=4803
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-CADDSuite-Docking-with-Ligand.pdf
Software documentation	http://www.ball-project.org/caddsuite
User documentation	http://www.ballview.org/Support/caddsuite-tutorial-1
Contact information	Dr. Jens Krüger e-mail: krueger@bioinformatik.uni-tuebingen.de Richard Grunzke e-mail: richard.grunzke@tu-dresden.de

Table 17, CADDSuite Docking with Ligand generation application technical details

8.9.6 CADDSuite: Docking without ligand generation

8.9.6.1Nature and Relevance

This workflow performs a docking procedure of ligands into a protein. This version allows more advanced user input and is thus suitable for users who have performed docking before. The actual docking process includes splitting of the receptor-file into protein and reference ligand, grid building, docking and rescoring of the dock results. For this workflow a protein structure file containing a protein in complex with a reference ligand and a file with the screening library (i.e., ligands that are to be docked into the protein) is required. The imported screening library will be checked, but no 3D coordinates will be created for the ligands - thus it is the user's responsibility to make sure the ligand input is correctly formatted (sdf MUST contain all hydrogen atoms of each molecule and all 3 coordinates for each atom).

8.9.6.2Usage

This workflow is used by logging in to the MoSGrid Science Gateway, then choosing the “Simulation” menu, “Docking”, “CADDSuite”, and the “Docking using CADDSuite using existing ligands” workflow. After choosing input files and parameters the workflow is to be submitted. After the workflow is finished the results are available for download.

8.9.6.3Software Details

The CADDSuite is a modular docking tool suite, offering various tools for the preparation, docking, scoring and analysis of protein ligand interactions. CADDSuite tools have been already used in Galaxy, gUSE, UNICORE and KNIME workflows. Using them up to multiple million ligands can be evaluated in short time.
It is available under http://www.ball-project.org/caddsuite; the usage is described under http://www.ballview.org/Support/caddsuite-tutorial-1, and can be cited via http://link.springer.com/article/10.1186%2F1758-2946-4-S1-O2

8.9.6.4Infrastructure Details

The MoSGrid Science Gateway is used to execute the workflow. It is based on Liferay, gUSE, and WS-PGRADE. It uses the underlying XtreemFS for storage and UNICORE to access connected clusters.

8.9.6.5Virtual Organization Details

The VO MoSGrid is used.

8.9.6.6Workflow Details

The input file formats are PDB and SDF and the output file formats are PDB and SDF. The input data sizes are between 1 and 10 MB, the output sizes are about 100 MB, the memory usage is between 8 and 16 GB, and the disk usage is low. The processing time is about 10 seconds per ligand. A graph can be found below.

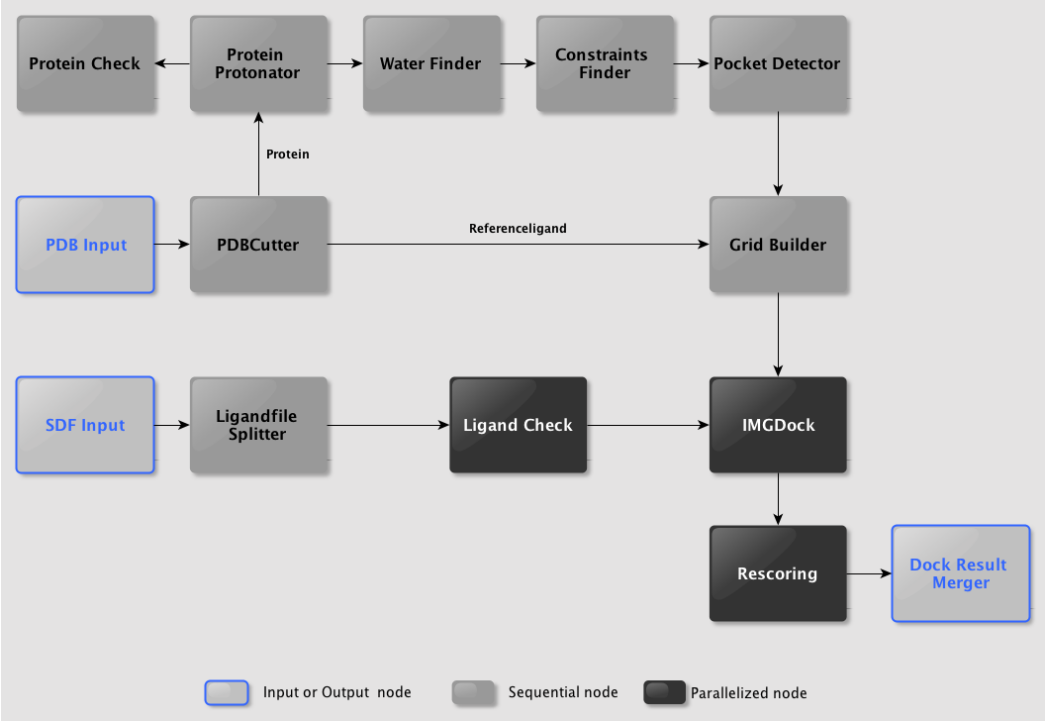

Figure 12, CADDSuite Docking without Ligand generation workflow (WS-PGRADE)

8.9.6.7 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=4802
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-CADDSuite-Docking-without-Ligand.pdf
Software documentation	http://www.ball-project.org/caddsuite
User documentation	http://www.ballview.org/Support/caddsuite-tutorial-1
Contact information	Dr. Jens Krüger e-mail: krueger@bioinformatik.uni-tuebingen.de Richard Grunzke e-mail: richard.grunzke@tu-dresden.de

Table 18, CADDSuite Docking without Ligand generation application technical details

8.9.7 NWChem: Basic workflow

8.9.7.1Nature and Relevance

This workflow performs a NWChem geometry optimisation of a desired molecule. It is highly useful for beginners who just want to perform a simple geometry optimisation but also as starting workflow for all upcoming type of jobs (Time-dependent DFT, solvent simulations etc.). The input file is a pre-prepared .nw file but for future applications we envision the direct usage of .xyz files as well.

8.9.7.2Usage

This workflow is used by logging in to the MoSGrid Science Gateway, then choosing the "Simulation" menu, "Quantum Chemistry", "NWChem", and the "Basic workflow" workflow. After choosing input files and parameters the workflow is to be submitted. After the workflow is finished the results are available for download.

8.9.7.3Software Details

NWChem aims to provide its users with computational chemistry tools that are scalable both in their ability to treat large scientific computational chemistry problems efficiently, and in their use of available parallel computing resources from high-performance parallel supercomputers to conventional workstation clusters.

NWChem software can handle

- Biomolecules, nanostructures, and solid-state
- From quantum to classical, and all combinations
- Ground and excited-states
- Gaussian basis functions or plane-waves
- Scaling from one to thousands of processors
- Properties and relativistic effects

NWChem is actively developed by a consortium of developers and maintained by the EMSL located at the Pacific Northwest National Laboratory (PNNL) in Washington State. Researchers interested in contributing to NWChem should review the Developers page.
The current version of NWChem is version 6.3 can be downloaded open-source under www.nwchem-sw.org/index.php/Download. The manual can be found under http://www.nwchem-sw.org/index.php/Release62:NWChem_Documentation. NWChem can be cited via http://www.sciencedirect.com/science/article/pii/S0010465510001438.

8.9.7.4Infrastructure Details

The MoSGrid Science Gateway is used to execute the workflow. It is based on Liferay, gUSE, and WS-PGRADE. It uses the underlying XtreemFS for storage and UNICORE to access connected clusters.

8.9.7.5Virtual Organization Details

The VO MoSGrid is used.

8.9.7.6Workflow Details

The input file format is NW and the output file formats are OUT but also HESS, ZMAT, CUBE and many more if desired for printing out additional output. The input data sizes are between 1 and 10 KB, the output sizes are about 1-100 MB, the memory usage is between 8 and 32 GB, and the disk usage is low. The processing time is between minutes and weeks

depending on the size of the molecule. The workflow is so fundamental and simple that a graph can be almost omitted here: one NW file serves as input, is submitted to the grid and the output comes back. More complicated workflows are based on this fundamental step (described in the following sections).

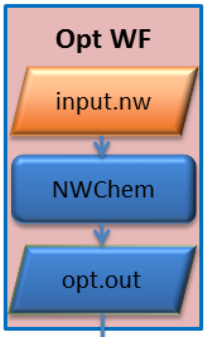

Figure 13, Basic NWChem workflow (WS-PGRADE)

8.9.7.7 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=3958
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-NWChem-Basic.pdf
Software documentation	www.nwchem-sw.org/index.php/Download
User documentation	http://www.nwchem-sw.org/index.php/Release62:NWChem_Documentation
Contact information	Dr. A. Hoffmann alexander.hoffmann@cup.uni-muenchen.de Mouala Moumin mouala.moumin@cup.uni-muenchen.de

Table 19, Basic NWChem application technical details

8.9.8 NWChem: Optimisation plus frequency calculation

8.9.8.1Nature and Relevance

This is a useful workflow because it combines two fundamental steps in quantum chemistry: a geometry optimisation and a frequency calculation which is needed to confirm the stability of the minimum.
The input file is a pre-prepared .nw file of a desired molecule. The geometry is parsed by a converter (bash shell script), combined with an “empty” .nw frequency file which is then submitted again. After this step, standard output of NWChem can be obtained.

8.9.8.2Usage

This workflow is used by logging in to the MoSGrid Science Gateway, then choosing the “Simulation” menu, “Quantum Chemistry”, “NWChem”, and the workflows “optfreq”. After choosing input files and parameters the workflow is to be submitted. After the workflow is finished the results are available for download.

8.9.8.3Software Details

NWChem aims to provide its users with computational chemistry tools that are scalable both in their ability to treat large scientific computational chemistry problems efficiently, and in their use of available parallel computing resources from high-performance parallel supercomputers to conventional workstation clusters.

NWChem software can handle

- Biomolecules, nanostructures, and solid-state
- From quantum to classical, and all combinations
- Ground and excited-states
- Gaussian basis functions or plane-waves
- Scaling from one to thousands of processors
- Properties and relativistic effects

NWChem is actively developed by a consortium of developers and maintained by the EMSL located at the Pacific Northwest National Laboratory (PNNL) in Washington State. Researchers interested in contributing to NWChem should review the Developers page.
The current version of NWChem is version 6.3 can be downloaded open-source under www.nwchem-sw.org/index.php/Download. The manual can be found under http://www.nwchem-sw.org/index.php/Release62:NWChem_Documentation. NWChem can be cited via http://www.sciencedirect.com/science/article/pii/S0010465510001438.

8.9.8.4Infrastructure Details

The MoSGrid Science Gateway is used to execute the workflow. It is based on Liferay, gUSE, and WS-PGRADE. It uses the underlying XtreemFS for storage and UNICORE to access connected clusters.

8.9.8.5Virtual Organization Details

The VO MoSGrid is used.

8.9.8.6Workflow Details

The input file format is OUT and the output file formats are OUT but also HESS, ZMAT, CUBE and many more if desired for printing out additional output. The input data sizes are between 1 and 10 KB, the output sizes are about 1-100 MB, the memory usage is between 8

and 32 GB, and the disk usage is low. The processing time is between minutes and weeks depending on the size of the molecule. Integrated into the workflow are “empty” .nw files which are combined with extracted geometry data to new .nw input files (Figure XXWF). The converter combines the extracted geometry with suited .nw files and gives these as true .nw input files for the corresponding jobs into the NWChem processing.

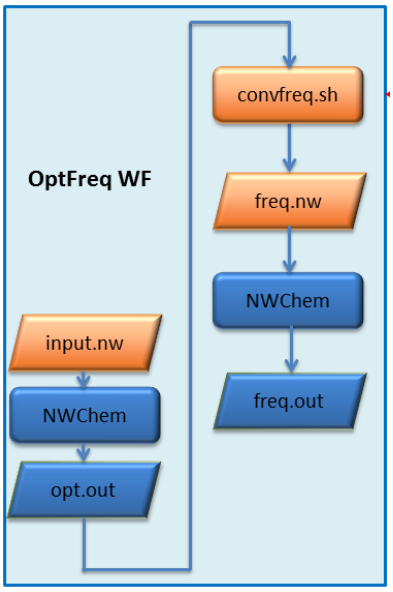

Figure 14, Workflow for NWChem optimisation and subsequent frequency calculation

8.9.8.7 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=3959
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-NWChem-Opt+freq.pdf
Software documentation	www.nwchem-sw.org/index.php/Download
User documentation	http://www.nwchem-sw.org/index.php/Release62:NWChem_Documentation
Contact information	Dr. A. Hoffmannalexander.hoffmann@cup.uni-muenchen.de Mouala Mouminmouala.moumin@cup.uni-muenchen.de

Table 20, NWChem optimisation and subsequent frequency calculation application technical details

8.9.9 NWChem: Several workflows after conversion

8.9.9.1Nature and Relevance

We have developed a series of workflows using NWChem based on the geometry-optimised structure. It is highly useful for beginners who just want to perform further calculation steps but also as building blocks for more complex workflows (meta-workflows, vide infra). The input file is a pre-prepared opt.out file of a preceding optimisation simulation. A frequency workflow (freq WF) performs a frequency calculation, a step which is always needed to characterise a minimum as stable minimum. A time-dependent DFT workflow (TD WF) performs a TD-DFT calculation in order to obtain the optical properties of a molecule. The Mulliken workflow (Mulliken WF) calculates the Mulliken charges of every atom of a molecule. The solvent workflow (Solvation WF) arranges a solvent sheath around a molecule and performs a further optimisation step. The similarity of these four workflows is that they are based on the output of the basic workflow which can be extracted in a conversion step (this step contains bash shell script).

8.9.9.2Usage

These workflows are used by logging in to the MoSGrid Science Gateway, then choosing the “Simulation” menu, “Quantum Chemistry”, “NWChem”, and the single workflows with the abbreviations **TD, Mull, Freq** and **Solv**. After choosing input files and parameters the workflow can be submitted. After the workflow is finished the results are available for download.

8.9.9.3Software Details

NWChem aims to provide its users with computational chemistry tools that are scalable both in their ability to treat large scientific computational chemistry problems efficiently, and in their use of available parallel computing resources from high-performance parallel supercomputers to conventional workstation clusters.

NWChem is actively developed by a consortium of developers and maintained by the EMSL located at the Pacific Northwest National Laboratory (PNNL) in Washington State. Researchers interested in contributing to NWChem should review the Developers page.
The current version of NWChem is version 6.3 can be downloaded open-source under www.nwchem-sw.org/index.php/Download. The manual can be found under http://www.nwchem-sw.org/index.php/Release62:NWChem_Documentation. NWChem can be cited via http://www.sciencedirect.com/science/article/pii/S0010465510001438.

8.9.9.4Infrastructure Details

The MoSGrid Science Gateway is used to execute the workflow. It is based on Liferay, gUSE, and WS-PGRADE. It uses the underlying XtreemFS for storage and UNICORE to access connected clusters.

8.9.9.5Virtual Organization Details

The VO MoSGrid is used.

8.9.9.6Workflow Details

In all four workflows the input file format is OUT and the output file formats are OUT but also HESS, ZMAT, CUBE and many more if desired for printing out additional output. The input data sizes are between 1 and 10 KB, the output sizes are about 1-100 MB, the memory usage is between 8 and 32 GB, and the disk usage is low. The processing time is between minutes and weeks depending on the size of the molecule. Integrated into the workflow are

"empty" .nw files which are combined with extracted geometry data to new .nw input files. The converter combines the extracted geometry with suited .nw files and gives these as true .nw input files for the corresponding jobs into the NWChem processing.

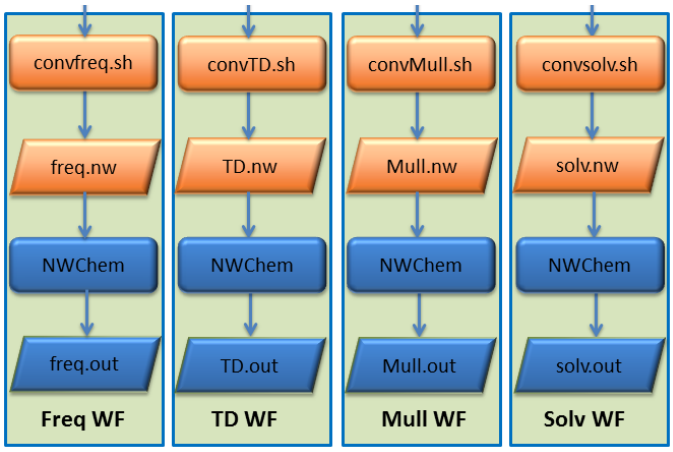

Figure 15, Several NWChem workflows which can follow up the basic workflow (WS-PGRADE)

8.9.9.7 Further Technical Details

Information	URL
SHIWA Repository Freq WF	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=4206
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-NWChem-Freq.pdf
SHIWA Repository TD-DFT WF	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=4751
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-NWChem-TD-DFT.pdf
SHIWA Repository Mulliken WF	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=4753
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-NWChem-Mulliken.pdf
SHIWA Repository Solvation WF	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=4752
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-NWChem-Solvation.pdf
Software documentation	www.nwchem-sw.org/index.php/Download
User documentation	http://www.nwchem-sw.org/index.php/Release62:NWChem_Documentation
Contact information	Dr. A. Hoffmann alexander.hoffmann@cup.uni-muenchen.de Mouala Moumin mouala.moumin@cup.uni-muenchen.de

Table 21, NWChem applications (Freq, TD, Mull, Solv) technical details

8.9.10 NWChem: Transition State search

8.9.10.1 Nature and Relevance

The search for transition states is a complicated process, but crucial in the theoretical description of chemical reaction mechanisms. Here, the computational challenge lies in the fact that the simulation code does not have to search for an energetic minimum, but for a saddle point (a higher-order minimum). This is time-consuming and often fails. If the user has really found such a saddle-point, it can only be verified by a following frequency calculation. This indicates by zero negative frequencies a normal minimum and by one negative frequency a saddle-point.
We have developed a rather simple workflow using NWChem which performs a transition state search. It is useful for more advanced users who want to combine it with some of the above-described sub-workflows to a more complex workflows (meta-workflows, vide infra). The input file is a pre-prepared TS.nw file.

8.9.10.2 Usage

This workflow is used by logging in to the MoSGrid Science Gateway, then choosing the “Simulation” menu, “Quantum Chemistry”, “NWChem”, and the single workflows with the abbreviations TS search. After choosing input files and parameters the workflow is to be submitted. After the workflow is finished the results are available for download.

8.9.10.3 Software Details

NWChem aims to provide its users with computational chemistry tools that are scalable both in their ability to treat large scientific computational chemistry problems efficiently, and in their use of available parallel computing resources from high-performance parallel supercomputers to conventional workstation clusters.

NWChem is actively developed by a consortium of developers and maintained by the EMSL located at the Pacific Northwest National Laboratory (PNNL) in Washington State. Researchers interested in contributing to NWChem should review the Developers page.
The current version of NWChem is version 6.3 can be downloaded open-source under www.nwchem-sw.org/index.php/Download. The manual can be found under http://www.nwchem-sw.org/index.php/Release62:NWChem_Documentation. NWChem can be cited via http://www.sciencedirect.com/science/article/pii/S0010465510001438.

8.9.10.4 Infrastructure Details

The MoSGrid Science Gateway is used to execute the workflow. It is based on Liferay, gUSE, and WS-PGRADE. It uses the underlying XtreemFS for storage and UNICORE to access connected clusters.

8.9.10.5 Virtual Organization Details

The VO MoSGrid is used.

8.9.10.6 Workflow Details

The input file format is OUT and the output file formats are OUT, but also HESS, ZMAT, CUBE and many more if desired for printing out additional output. The input data sizes are between 1 and 10 KB, the output sizes are about 1-100 MB, the memory usage is between 8 and 32 GB, and the disk usage is low. The processing time is between minutes and weeks depending on the size of the molecule. This workflow strongly resembles the basic NWChem workflow in Figure 13, since it only uses a TS.nw file which contains the molecules´ coordinates, information about functional, basis set, spin, multiplicity and – most important –

the desired atoms which should approach each other. This nw-file is more complicated than the other ones described herein. Hence, it is given in the following:

```
title "parent diels-alder - constrained geometry optimization + saddle search"

# norbornene = -45.69811295 au  r = 1.539 Angstroms

geometry autosym # Equil nonbornene geometry
C       0.48438728   -1.26420666    0.66379632
C       0.48438728   -1.26420666   -0.66379632
C      -0.32763055   -0.09123377    1.11105174
C      -0.32763055   -0.09123377   -1.11105174
C      -1.36089042   -0.02893165    0.00000000
C       0.50866739    1.15402721    0.76530727
C       0.50866739    1.15402721   -0.76530727
H      -0.69016477   -0.11291256   -2.13538649
H      -0.69016477   -0.11291256    2.13538649
H      -1.93888714    0.89733827    0.00000000
H       1.50914212    1.10998500   -1.19387233
H       1.50914212    1.10998500    1.19387233
H       1.09450829   -1.87707752   -1.31279758
H       1.09450829   -1.87707752    1.31279758
H      -2.02362675   -0.89310262    0.00000000
H       0.02217904    2.04804733    1.15788544
H       0.02217904    2.04804733   -1.15788544
end
driver
  maxiter 101
end

basis
* library 6-31G*
end

dft
  xc b3lyp
end

geometry adjust #move to guess geometry & apply constraints
  zcoord
    bond 6 3 2.267 r constant
    bond 7 4 2.267 r constant
  end
end
task dft optimize #relax with constraints

geometry adjust #release constraints
  zcoord
    bond 6 3 2.267 r
    bond 7 4 2.267 r
  end
end
task dft saddle #go for the transiton state
```

8.9.10.7 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/details-view.xhtml?appid=4913
Application description template	http://www.erflow.eu/documents/388575/771342/Application-description-NWChem-Transition-State.pdf
Software documentation	www.nwchem-sw.org/index.php/Download
User documentation	http://www.nwchem-sw.org/index.php/Release62:NWChem_Documentation
Contact information	Dr. A. Hoffmann alexander.hoffmann@cup.uni-muenchen.de Mouala Moumin mouala.moumin@cup.uni-muenchen.de

Table 22, NWChem transition state search application technical details

8.10 Porting Experience

The porting of the "Energy Minimisation", "Equilibration", and "AutodockVinaFull" went flawlessly. The porting of "Docking with ligand generation" and "Docking without ligand generation" was at first prevented by an error message ("Publish forbidden"). With "Single TPR" it was at first prevented by gUse bug 151. Each problem was promptly investigated by the SHIWA team and fixed.
The porting of the simple NWChem workflows went without large problems. The SHIWA team helped very competently and efficiently. The more complex workflows such as NWChem-Opt+freq could be ported as well, but the effort on chemist´s side was larger.
The really complex meta-workflows such as the spectroscopic workflow have not been successfully ported yet. However, with all fundamental NWChem workflows, we have built up a very useful toolbox to combine helpful meta-workflows for quantum chemists.

9 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.
The applications considered during the first year of the ER-FLOW project have all been developed in the **HELIO project**[7], an FP7 project that has developed a Virtual Observatory for Heliophysics based on a distributed, service oriented architecture. Two key components of HELIO are a set of web services that cover important functionalities for the heliophysicst and a workflow engine (**TAVERNA**)[8] used to orchestrate and combine the services.

9.1 Ported Applications

All the 6 workflows ported during the first year of the project have been developed within the HELIO project. To understand the nature and utility of these workflows it is advantageous to briefly describe some common investigation steps in the discipline.

- **Find what is interesting**: During this step, the scientist browses catalogues of events that were observed throughout the solar system and catalogues of features that were observed on the Sun's surface to establish what is worth investigating
- **Find which observations are available**: During this step the scientists browses catalogues of instruments capabilities and locations to understand which observations are available to investigate the events or features of interest.
- **Find the data**: During this step the scientists find and access the data sets that are available and relevant to study the events and features.
- **Analyse the data:** During this step the scientists analyse the retrieved data sets.

All the ported workflows cover all or part of this process to investigate events either in a single point of the Solar System or to investigate how events propagate throughout the Solar System by means of a Propagation Model. The ported applications cover one or more of the listed steps. The ported applications fall into two broad categories: workflows of general interest and workflows specific to certain events.

Workflows of general interest:

- Overview of events over a given period, used by scientists to assess whether a period is interesting or not.
- Synoptical map of the Surface of the Sun, used by scientists to assess the presence of features on the surface of the sun for a given date.
- Capability of retrieving all the data available for a certain event.

Workflows to investigate specific events. The workflows that investigate specific phenomena cover Coronal Mass Ejection, Solar Wind fluctuations and Solar Flares, which are among the most relevant events studied in Heliophysics

[7] http://www.helio-vo.eu/index.php
[8] http://www.taverna.org.uk/

Workflow	Description
Fast CME Propagation	Models the propagation of the fastest Coronal Mass Ejections (CMEs) throughout the Solar System
Flare Events Data	Returns all the data available for the study of Solar Flares
Monthly Event Overview	Returns an overview of all the relevant events that occurred each month within the selected period
Generic Event Data	Retrieves all available data of a given event
Origins of the Solar Wind	Correlates features on the Sun to Solar Wind fluctuations detected at Earth
Solar Surface Map	Returns the synoptical map of the surface of Sun for a given day.
Contact Detail	pierang@cs.tcd.ie

Table 23, Heliophysics Workflows Overview

9.2 Applications Usage

All the ported workflows are used by the users in the same way. There are three possible access modalities that are being developed in the project:

- Through the **SHIWA Simulation Platform (SSP)**. This modality is preferred by developers during the porting and testing phase.
- Through the User Interface. This modality is preferred for the scientists to test the applications and it is the first step for all applications.
- Through specifically developed **Application Specific Module (ASM)** that will be developed within the SCI-BUS project for the workflows that are most useful and popular in the Heliophysics community.

9.3 Technical Background

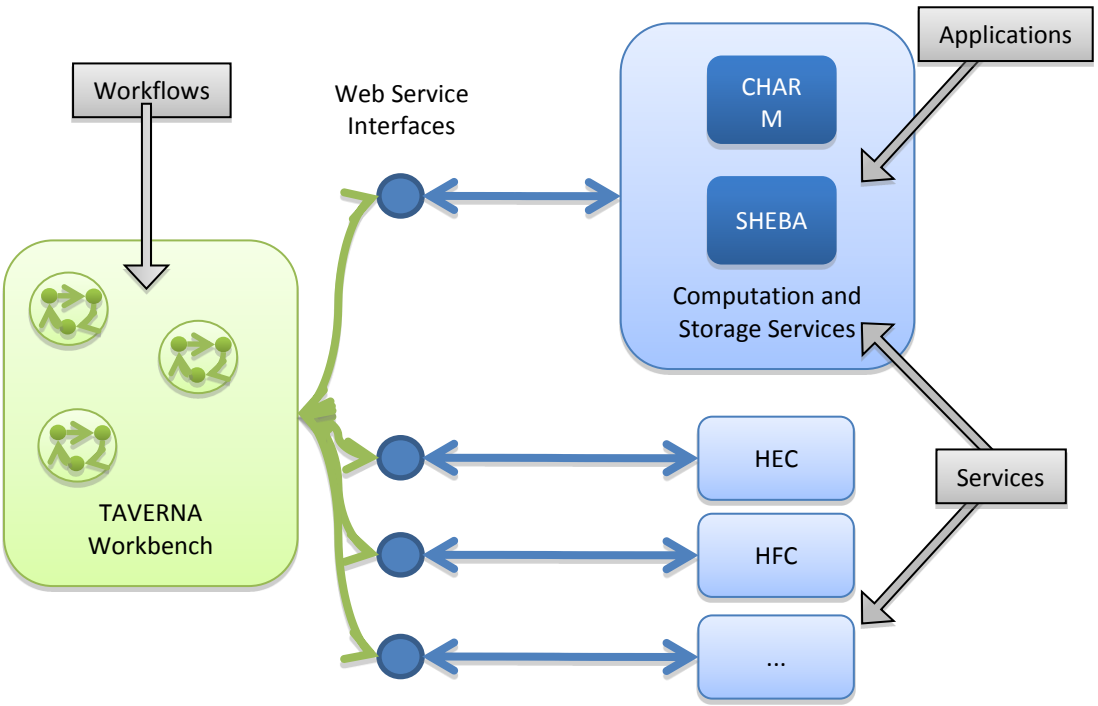

Figure 16, Services, Workflows and Application of HELIO

The six ported 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 image 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 first 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.

9.4 Applications Description

9.4.1 Fast CME Propagation

9.4.1.1 Nature and Relevance

This workflow is used to investigate one of the most relevant events in HELIOphysics; Coronal Mass Ejections[9]. The propagation of these events is studied throughout the Solar System by the means of the execution of the SHEBA propagation model on the HELIO Processing Service. This workflow was chosen for porting because the study of propagation of events is a key process in Heliophysics and because CMEs are among the most relevant events to be analysed in the field. The workflows studies only the fastest CME assuming that this kind of event are less influenced by the drag caused by the magnetic field of the Solar System.

9.4.1.2 Usage

To use this workflow the user has only to input the time range to be analysed, the output consists in the Expected Arrival Times (ETA) for each of the bodies of the Solar System where there are instruments.

9.4.1.3 Software Details

This workflow uses the HELIO services listed in the table below.

Service	Usage in this workflow
HEC	Return events of a given kind for a time period
HPS	Executes the **SHEBA** propagation model

Table 24, HELIO Services used by workflow that studies the propagation of fast CMEs

9.4.1.4 Workflow Details

The modified workflow is described in Figure 17.

9.4.1.5 Further Technical Details

Information	URL
Original TAVERNA workflow	http://www.myexperiment.org/workflows/3262.html
Modified TAVERNA workflow	http://www.myexperiment.org/workflows/3469.html
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4155
ER-FLOW technical templates	http://www.erflow.eu/documents/388575/775508/helio.application.2.FastCME.pdf
Other user documentation	http://www.helio-vo.eu/services/service_interfaces.php
Contact Detail	pierang@cs.tcd.ie

Table 25, Information sources for the workflow that studies the propagation of the fastest CMEs

[9] http://en.wikipedia.org/wiki/Coronal_mass_ejection

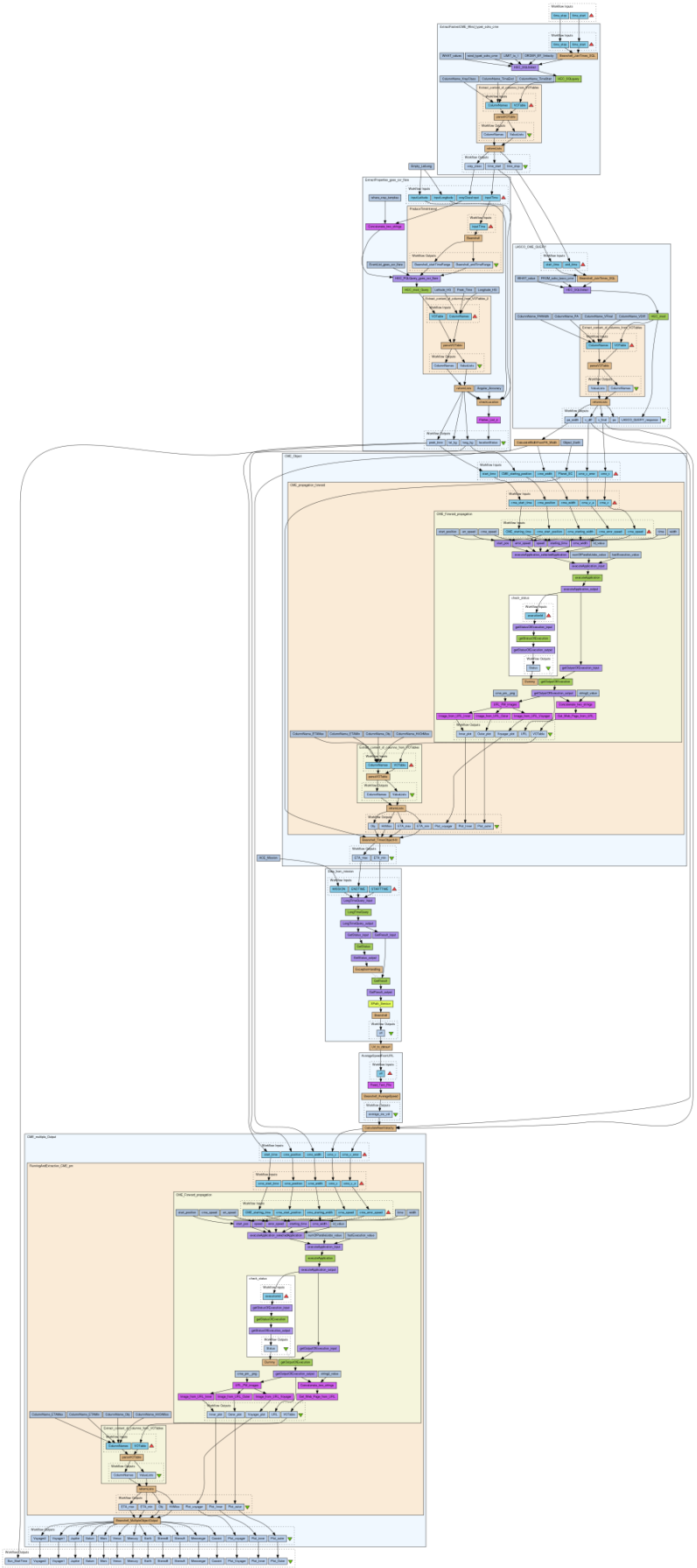

Figure 17, Workflow to investigate the propagation of the Fastest CMEs (TAVERNA)

9.4.2 Find Data for Flare Events

9.4.2.1 Nature and Relevance

This workflow returns all the data available for the specified instruments that are relevant to the study of Solar Flares[10] of the specified kind.

Given a time range, this workflow looks for flares within the specified energy range (GOES x-ray flare class), and provides the observations for such time range for the list of instruments asked. It also provides the table of flares with its properties.

Solar Flares are very important events at the centre of investigation in Heliophysics; this workflow is relevant to the community as it allows accessing all data available for the specified instruments relevant to the flares within specified energy range

9.4.2.2 Usage

To use this workflow the user has to input the time range to be analysed, the type of flares in investigation and the list of instruments of interest. The output consists in the list of all relevant available data and a table with the characteristics of the Flares being investigated.

9.4.2.3 Software Details

This workflow uses the HELIO services listed in the table below.

Service	Usage in this workflow
HEC	Returns flares of a given energy for a time period

Table 26, HELIO Services used by workflow that finds all data for Flare Events

9.4.2.4 Workflow Details

The modified workflow is described in Figure 18.

9.4.2.5 Further Technical Details

Information	URL
Original TAVERNA workflow	http://www.myexperiment.org/workflows/940.html
Modified TAVERNA workflow	http://www.myexperiment.org/workflows/3467.html
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4155
ER-FLOW technical templates	http://www.erflow.eu/documents/388575/775508/helio.application.4.DataForFlareEvents.pdf
Other user documentation	http://www.helio-vo.eu/services/service_interfaces.php
Contact Detail	pierang@cs.tcd.ie

Table 27, Information sources for the workflow that finds all data for Flare Events

[10] http://en.wikipedia.org/wiki/Solar_flare

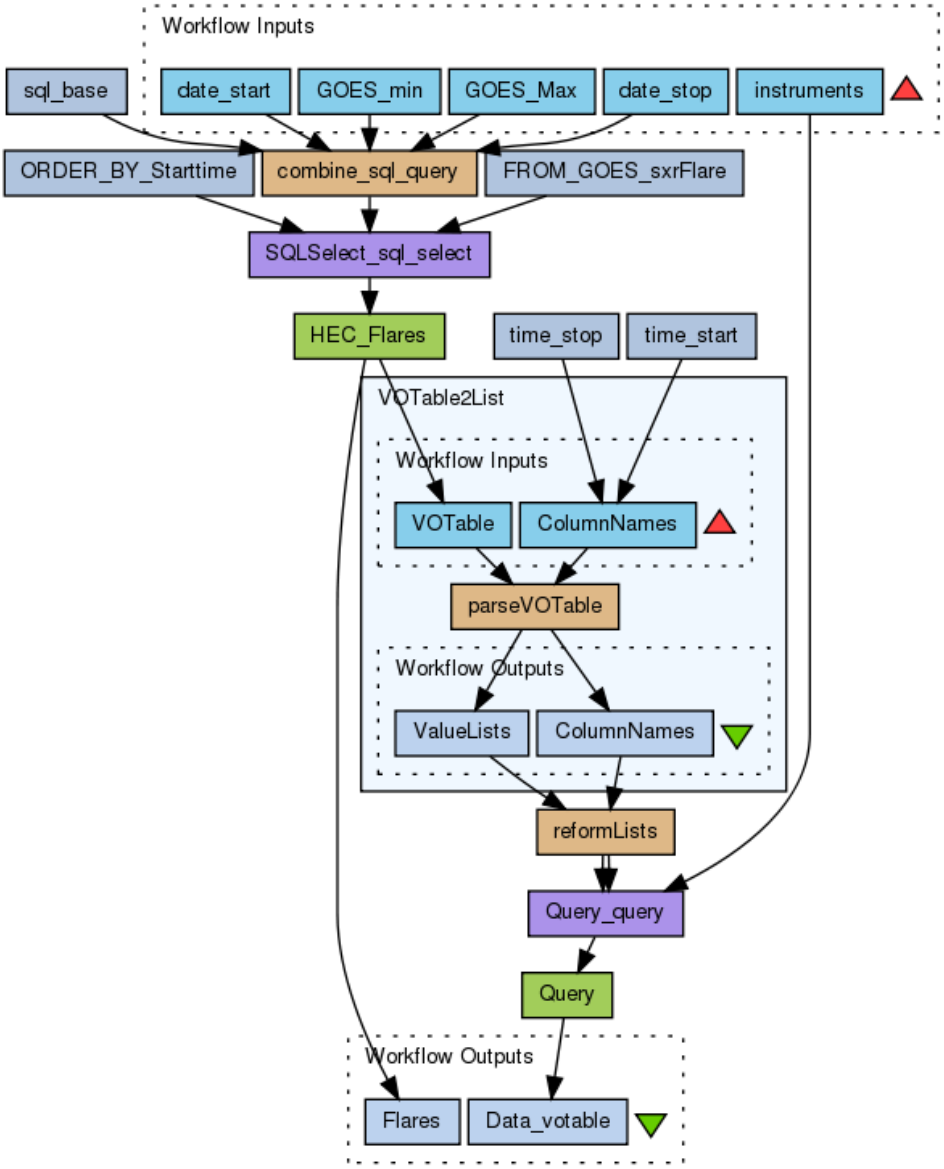

Figure 18, Workflow that finds all data for Flare Events (TAVERNA)

9.4.3 Overview of Events for the month.

9.4.3.1 Nature and Relevance

This workflow returns an overview of all the relevant events that occurred each month within the selected period. It is highly relevant to the community as it presents the uses with a recapitulation of events and the presence of certain events is quite often the first step of an investigation for a Heliophysicist. An overview of all the events spread during a period allows the scientist to understand which periods can be interesting to be investigated further with other workflows.

9.4.3.2 Usage

To use this workflow the user has to input the time range to be analysed and the type of event under investigation (by specifying the table to be used). The output consists in the list of all relevant events available as a VOTable and as a list of lists.

9.4.3.3 Software Details

This workflow uses the HELIO services listed in the table below.

Service	Usage in this workflow
HEC	Queries the specified table and returns a list of all relevant events given a time range.

Table 28, HELIO Services used by workflow that finds all events in a given time range

9.4.3.4 Workflow Details

The modified workflow is described in Figure 19.

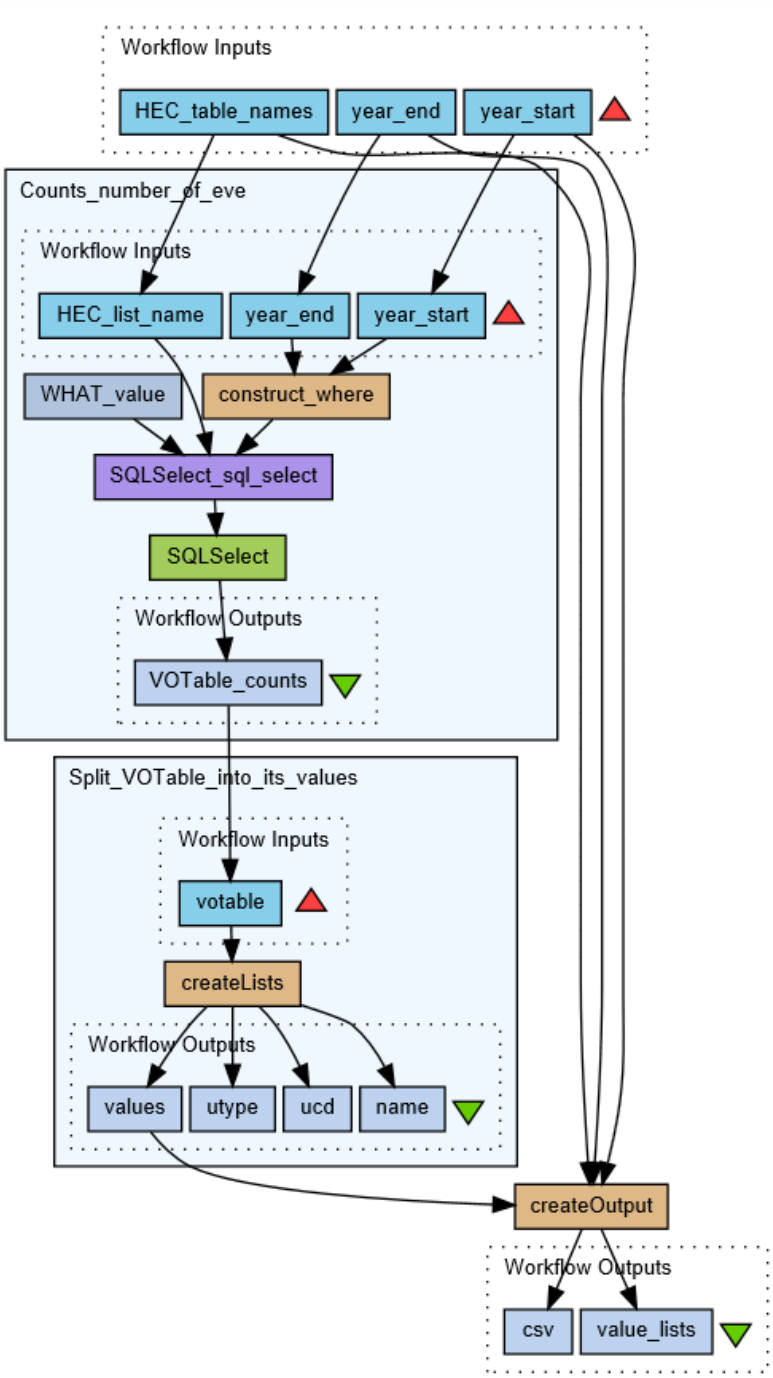

Figure 19, Workflow to obtain a list of relevant events during a given time period (TAVERNA)

9.4.3.5 Further Technical Details

Information	URL
Original TAVERNA workflow	http://www.myexperiment.org/workflows/3119.html
Modified TAVERNA workflow	http://www.myexperiment.org/workflows/3468.html
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=3945
ER-FLOW technical templates	http://www.erflow.eu/documents/388575/775508/helio.application.5.EventsOverview.pdf
Other user documentation	http://www.helio-vo.eu/services/service_interfaces.php
Contact Detail	pierang@cs.tcd.ie

Table 29, Information sources for the workflow that returns a summary of relevant events for a given time period

9.4.4 Retrieval of all available data for a given event.

9.4.4.1 Nature and Relevance

This workflow is used to retrieve all available data that is available for a certain event; it is a natural completion to the one described in section 9.4.3. After relevant events are found, data is retrieved for further investigation. Data sources are selected by defining the capabilities, i.e. the type of observations performed by the various instruments. In order to obtain this data, three different data sources are queried, the **HELIO Event Catalogue** to retrieve information on the events and the **Instrument Capability Service** that returns information on the observational capabilities of the various instruments and the **Data Access Provider Service,** a centralized service that access different data sources**.**

9.4.4.2 Usage

To use this workflow the user has to input the time range to be analysed and the type of event under investigation (by specifying the table of the HELIO Event Catalogue to be used). The output consists in two VO Tables: That returned by the HELIO Event Catalogue with all the information describing the event and that returned by the Data Provider Access Service with all the urls with the available data.

9.4.4.3 Software Details

This workflow uses the HELIO services listed in Table 30

Service	Usage in this workflow
HEC	Queries the specified table and returns a list of all relevant events given a time range.
ICS	Returns the observational capabilities of an instrument
DPAS	Centralized service that returns the urls for available data.

Table 30, HELIO Services used by workflow that finds all data relevant to one event.

9.4.4.4 Workflow Details

The modified workflow is described in Figure 20.

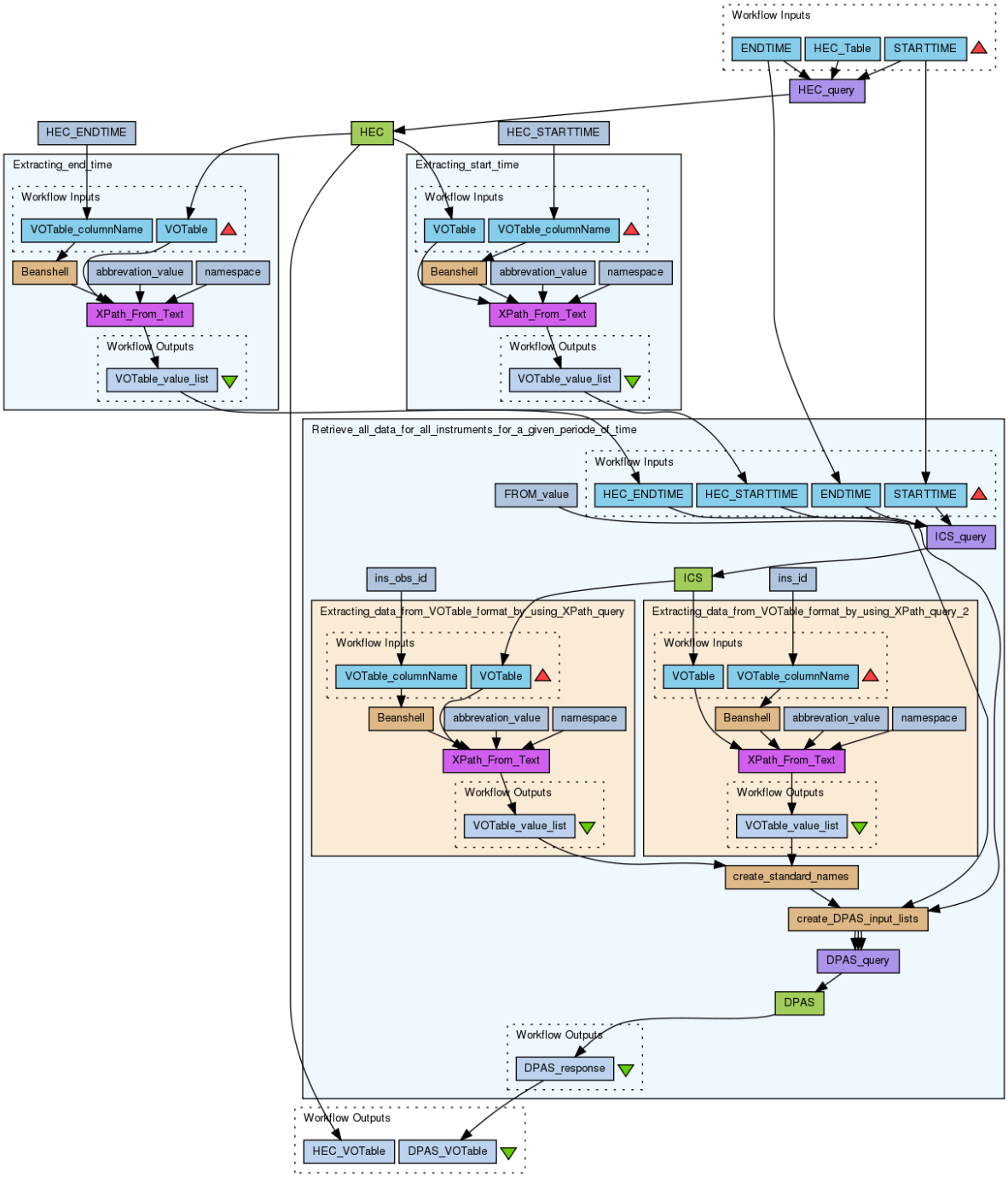

Figure 20, Workflow to obtain all data pertaining a given event (TAVERNA)

9.4.4.5 Further Technical Details

Information	URL
Original TAVERNA workflow	http://www.myexperiment.org/workflows/3119.html
Modified TAVERNA workflow	http://www.myexperiment.org/workflows/3468.html
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=3960
ER-FLOW technical templates	http://www.erflow.eu/documents/388575/775508/helio.application.3.EventData.pdf
Other user documentation	http://www.helio-vo.eu/services/service_interfaces.php
Contact Detail	pierang@cs.tcd.ie

Table 31, Information sources for the workflow that returns a summary of relevant events for a given time period

9.4.5 Finds the origins on the Sun of the Solar Wind

9.4.5.1 Nature and Relevance

This workflow correlates features on the Sun to Solar Wind fluctuations detected at Earth.
This is another workflow that deals with specific phenomena like those described in 9.4.1 and 0 rather that generic analysis steps such as 9.4.3 or 9.4.4. In this case, it tries to find features on the surface of the Sun that are plausible causes for fluctuations in the solar wind detected at Earth. Solar Wind is a physical phenomenon that is quite significant for Heliophysicists and investigating the relationship between perceived fluctuations at Earth and the features on the Solar Surface is a common scientific undertaking in the field.

9.4.5.2 Usage

To use this workflow the user has to input the time range (start and end time) and the type of event that has be analysed (Coronal Mass Ejection - CME or Co-rotating Interactive Regions - CIR). The output consists of five different products: the plot of Solar Wind parameters of one day around the requested period, a VOTable from either the HELIO Feature Catalogue or the HELIO Event Catalogue with the results of the Coronal Holes and Coronal Mass Ejection events catalogued, URL for either features map from HELIO Feature Catalogue or a Coronal Mass Ejection movie, and, finally, the URLs to the diagrams produced by the propagation model running on the HELIO Processing Service.

9.4.5.3 Software Details

This workflow uses the HELIO services listed in the table below.

Service	Usage in this workflow
HEC	Queries the specified table and returns a list of all relevant events given a time range.
HFC	Returns a map with all the features present on the Solar surface at the time requested by the user
HPS	Generic asynchronous processing service.
SHEBA	Propagation model that predicts when certain events reach planets and other bodies of the Solar System

Table 32, HELIO Services used by workflow that finds origins of the Solar Wind

9.4.5.4 Workflows Details

The modified workflow is described in Figure 21.

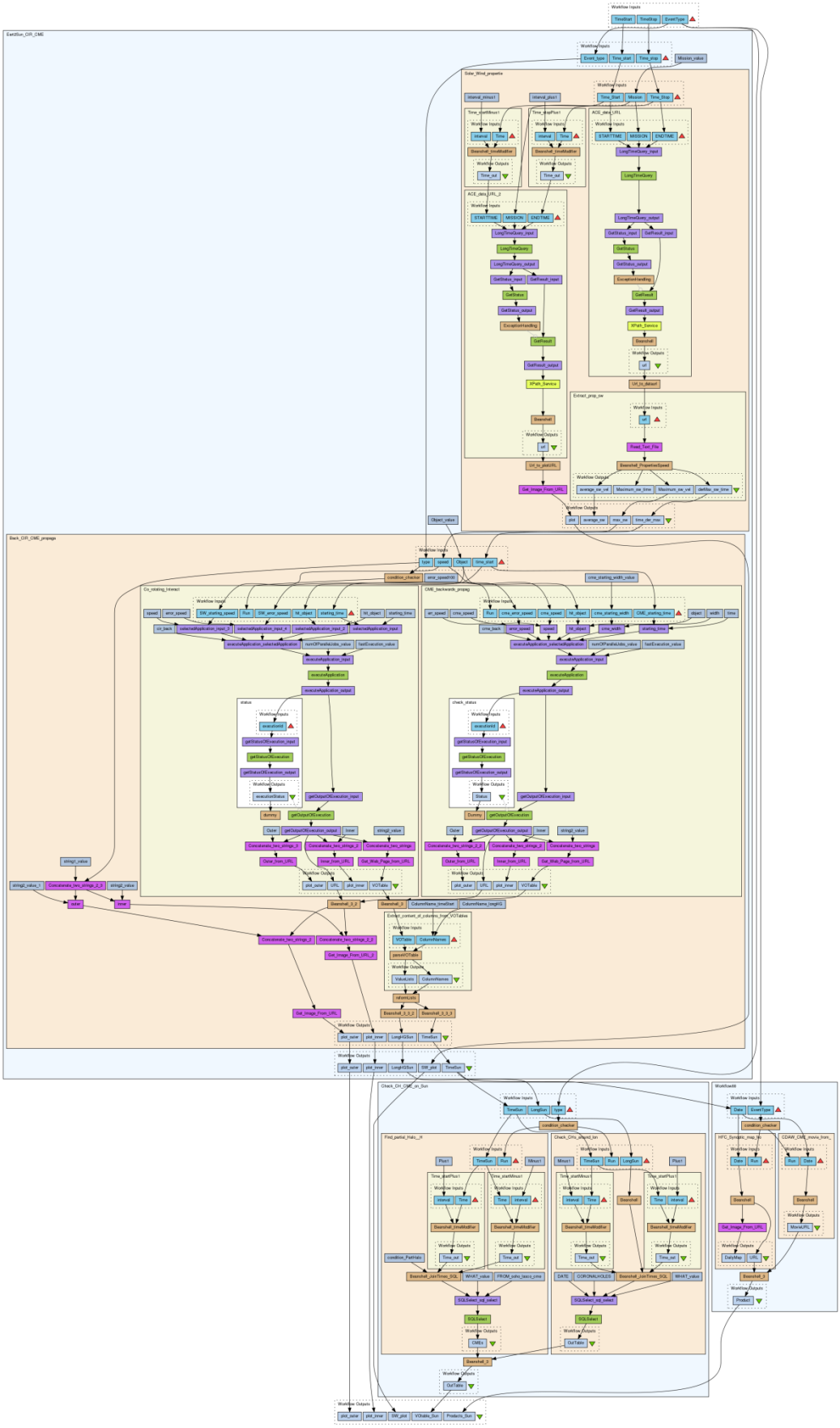

Figure 21, Workflow that finds the origins of the Solar Wind (TAVERNA)

9.4.5.5 Further Technical Details

Information	URL
TAVERNA workflow	http://www.myexperiment.org/workflows/3301.html
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=3956
ER-FLOW technical templates	http://www.erflow.eu/documents/388575/775508/helio.application.1.SolarWind.pdf
Other user documentation	http://www.helio-vo.eu/services/service_interfaces.php
Contact Detail	pierang@cs.tcd.ie

Table 33, Information sources for the workflow that returns a summary of relevant events for a given time period

9.4.6 Synoptical Map of the Features of the Surface of the Sun

9.4.6.1 Nature and Relevance

Returns the synoptical map of the surface of Sun for a given day. This workflow is generic, as it does not deal with any specific event or feature but rather return a synoptical map with all the relevant features of the surface of the sun. It is highly relevant to the community as features on the surface of the Sun are often related, if not directly the cause of, events that propagate throughout the Solar System and investigating such relationship is a common undertaking by the Heliophysicist.

9.4.6.2 Usage

The user enters the date he is interested in and the workflow return a map with all the features of the Solar surface.

9.4.6.3 Software Details

This workflow uses the HELIO services listed in the table below.

Service	Usage in this workflow
HFC	Queries catalogues of Features of the Solar Surface

Table 34, HELIO Service used by the workflow that returns the synoptical map of the Surface of the Sun

9.4.6.4 Workflow Details

The modified workflow is described in Figure 20.

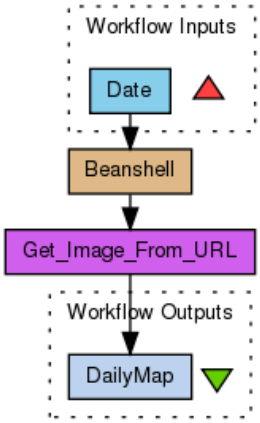

Figure 22, Workflow that returns the synoptical table of the surface of the Sun

9.4.6.5 Further Technical Details

Information	URL
TAVERNA workflow	http://www.myexperiment.org/workflows/3293.html
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4854
ER-FLOW technical templates	http://www.erflow.eu/documents/388575/775508/helio.application.6.SolarSurfaceMap.pdf
Other user documentation	http://www.helio-vo.eu/services/service_interfaces.php
Contact Detail	pierang@cs.tcd.ie

Table 35, Information sources for the workflow that returns a summary of relevant events for a given time period

9.5 Porting Experience

The process of porting of the TAVERNA workflows is divided in the following steps:

- Testing of the existing TAVERNA workflows,
- Porting of the workflows on the SHIWA Workflow Repository,
- Execution of the workflows on the SHIWA Simulation Platform.

During the first phase one problem has been found. As the workflows for heliophysics rely on the invocation of web services, when any of them is moved to another URL, the workflows becomes unusable. To address this issue, a two-stage approach has been adopted:

1. The existing workflows have been modified to invoke the services with the new URLs, and these new workflows have been saved in the myExperiment platform.
2. A TAVERNA plugin that invokes the HELIO Registry Services (HRS) to find the current deployment of each service has been developed at the Mullard Space Laboratory (MSSL) of UCL.

Whilst modifying the workflows to update the URLs of services, some modifications have been necessary to cope with changes in some of the services interfaces. Most workflows have been modified and re-deployed on myExperiment. A list of these updates can be found in Table 36.

Workflow name	Original myExperiment ID	New myExperiment ID
Fastest CME Propagation	3262	3469
Data about Flares	940	3467
All data for an event	1512	3466
Event counts in a time range	3119	3468

Table 36, Modified TAVERNA workflows in myExperiment

During the second phase another problem has been faced. TAVERNA workflows save results in files or folders depending on their type, more specifically lists are saved as folders containing files or more folders. WS-PGRADE, the main workflow engine that executes the different workflows as nodes of a meta workflow handles only files. To address this issue we have investigated to use the ***–outputfile*** option of the TAVERNA workflow engine that saves all the output as a single file. The result can then be either re-transformed in the original folder structure or directly investigated by means of the DATAVIEWER offered by the TAVERNA project.

10 Life Science

Life science (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, which is a small representation considering the size of the Life Science domain. 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 first project year we defined the use cases and the applications to port to SHIWA, developed new workflows using WS-PGRADE, and ported them to the SHIWA platform. We also connected the AMC science gateway under development at the SCI-BUS project (https://gateway.ebioscience.amc.nl) to the SHIWA repository to facilitate import and export of workflows. In Deliverable D5.1, AMC had set up some goals with regard to the usage and exploitation of the potentials of the SHIWA platform. We have moved towards these goals; below we give a brief explanation of how we have applied the envisioned strategies.

With respect to migrating workflows to WS-PGRADE, we have been quite successful with the porting of 11 applications and 30+ workflows. Some workflows already existed for MOTEUR, and have been revised and ported to the SHIWA platform, especially in the domains of neurosciences and NGS. Others have been developed in WS-PGRADE, some of them from scratch and others based on previous implementations in MOTEUR. Then the workflows were ported to the SHIWA platform. Additionally, we have added metadata to the workflows that are stored in the SHIWA repository; thus, this repository serves not only for sharing, but also as a documentation source for our workflows. We have furthermore used the SHIWA repository whenever we needed to share the workflows, which have made the sharing experience very smooth. For example, in our publications, the workflows are referred to by their ID's on the SHIWA repository. These ID's can be used in a standardized URL to uniquely point the browser to the description and implementations of that workflow.

10.1 Ported Applications

Given the variety of scientific domains of the biomedical problems addressed at the AMC, large scope of applications has been considered for porting as WS-PGRADE workflows to the SHIWA platform. The chosen applications display the highest requirements regarding data processing capacity; therefore, they are good candidates for porting to DCIs.

Applications in four scientific areas in which the AMC has largest expertise were selected for the first project year:

- Neuroimaging, mainly Magnetic Resonance Imaging (MRI) for structural and Diffusion Tensor Imaging (DTI). Applications include
  - Freesurfer, for brain segmentation in structural MRI scans;
  - DTI preprocessing, for preparation of DTI data for further processing;
  - FSL BedpostX, for tracking of white matter fibers in DTI scans; and
  - DTI Population Registration, for calculation of an average brain from DTI scans.
- Next generation sequencing (NGS) of humans. Applications include
  - Detection and annotation of Single Nucleotide Polymophisms (SNP);
  - Genome re-sequencing;
  - Sequence alignment, resampling and quality control;
- Biostatistics, in particular statistical modelling for various types of data (e.g. mass-spectrometry) using double cross validation techniques with various models:
  - Penalized Logistic Regression
  - Support Vector Machines (SVM)
  - Penalized Nonlinear Canonical Correlation Analysis (CCA)
- Molecular docking, in particular
  - AutoDock Vina, for virtual screening small molecules that can interact with proteins and modulate activity.

The main motivation for porting these applications to distributed infrastructures is the need to increase throughput for large data collections, which require their execution for different input data. Some applications also require manipulation of large files, or require long execution times (>24h per input dataset), therefore they have become impractical on the regular infrastructure available for researchers at the AMC.

An overview of the ported application is presented in the table below. The corresponding application templates (available on http://www.erflow.eu/applications) describe the basic information on each application and the expected input/output formats.

Domain	Application	ID in Shiwa Repo		DCI	WFMS
		Wf	Impl.		
Neurosciences	Freesurfer	2251	2552	Grid	MOTEUR
		4251	4600	Grid	WS-PGRADE
	DTI Preprocessing	1751	2001 2006*	Grid	MOTEUR
		4250	4599	Grid	WS-PGRADE
	BedpostX	1512	3208 1770*	Grid	MOTEUR
		4858	5109	Grid	WS-PGRADE
	DTI Population Registration	4601	4901	Grid	MOTEUR
BioStatistics	DCV	4805	5053	Grid	WS-PGRADE
Docking	Autodock Vina	4256	4605	Grid	WS-PGRADE
Next Generation Sequencing	Sequence Alignment (BWA)	1513	1772	Grid, Cluster	MOTEUR
	SNP Calling	4255	4604	Grid, Cluster	WS-PGRADE
	SNP Annotation	4254	4603	Grid, Cluster	WS-PGRADE
	DownSample	4253	4602	Grid	WS-PGRADE
	Sequence Assembly				

Table 37 - Overview of Life Sciences applications ported to the SHIWA platform. (*) These implementations are configured to run on SHIWA VO, whereas the rest run as VLEMED VO.

10.2 Applications Usage

The applications are used to process datasets owned by the biomedical researchers. Normally, the biomedical researcher starts the workflow from the customized interface of the science gateway. In this case, users 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.

The e-bioscience team of the AMC currently operates two science gateways that we coin "e-bioinfra gateway" and "AMC SCI-BUS gateway" in this text. The first is based on MOTEUR WfMS, and contains web applications implementing custom interfaces for selected workflows. The second is based on the platform provided by the SCI-BUS project (WS-PGRADE and LifeRay); currently, custom interfaces are available only for the neuroimaging applications. The characteristics of these gateways are summarized in the table below.

	e-bioinfra gateway	AMC SCI-BUS gateway
WfMS	MOTEUR	WS-PGRADE
DCI	Dutch Grid	Dutch Grid, local cluster
url gateway	http://www.ebioscience.amc.nl/ebioinfragateway/	http://gateway.ebioscience.amc.nl/
Documentation	http://bioinformaticslaboratory.nl/twiki/bin/view/EBioScience/EBioInfraGateUserDoc	http://bioinformaticslaboratory.nl/twiki/bin/view/EBioScience/WspgradeUserDoc
Applications (Sep. 2013)	Customized interfaces for Freesurfer, FSL BedPostX, DTI pre-processing, DTI Atlas, Double Cross Validation	All applications are available, but only the neuroscience applications have customized interface.

Table 38, Characteristics of science gateways available to run Life Science applications.

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.

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.

10.3 Technical Background

A thorough analysis of the characteristics of these workflows is presented in the paper "Understanding workflows for distributed computing[11]: nitty-gritty details", recently accepted to the *8th Workshop on Workflows in Support of Large-Scale Science (WORKS'2013)*, which will be held in conjunction with Supercomputing 2013, in November, Denver, US. The applications of interest at the AMC process a large variety of data types (images, DNA sequences, biostatistics data for example from mass spectrometry, and ligands data). They display different characteristics concerning data sizes, computing times, number of tasks, and data and control-flow patterns. Below we summarize the most interesting facts, from two perspectives: application/workflow characteristics and the underlying infrastructure characteristics.

[11] The paper is available to the reviewers and ER-flow members upon request before its publication in the proceedings.

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 WfMS, 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 WfMS on different data and parameter settings.

The distributed infrastructure used to run these 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 of the AMC network, offering a more reserved environment for privacy sensitive applications.

10.3.1 General information about NGS workflows

In case of NGS workflows, the developer and the end-user are the same. Therefore, the usage scenario of these workflows differs from the gateway itself. Also, these workflows need to run both on grid and on cluster resources. These characteristics of the NGS workflows result in special requirements for adequate solutions for data handling and parameter sweeps. The solutions explained below hold only for the NGS applications ported to WS-PGRADE; these explanations are omitted in the individual explanations of these applications in the remainder of this chapter (sections 10.4.7 to 10.4.11).

10.3.1.1 Data handling

DNA sequence data is nowadays in the order of several gigabytes per experiment. A typical exome[12] sample uses around 16 GB of storage space at the time of writing. In practice this means that the data transfer of these files should be minimized to reduce overhead, and that the data should not be transferred via the server where the gateway is hosted to prevent a bottleneck. To resolve the second issue, only file paths (references), and not the files themselves, are passed on from one component to the next in sequence analysis workflows. This feature is supported by WS-PGRADE only for gLite middleware, but not for PBS; therefore, a customized solution was designed for the NSG workflows.

The inputs and outputs of the workflows are plain text files that contain the complete path to the actual input and output files on the storage system. The main executable of each workflow component calls a script that handles the download of input files and uploads the output files. This is in fact a customized job wrapper, in addition to the wrapper generated automatically by WS-PGRADE. Currently, our solution supports file transfers for gLite, clusters and local systems. We expect very soon that this kind of data transfer across different available DCI’s will be supported directly by WS-PGRADE, after which the implementation of these workflows will also be simplified.

[12] https://en.wikipedia.org/wiki/Exome_Sequencing

10.3.1.2 Parameter sweeps

WS-PGRADE supports parameter sweeps. However, files need to have a specific name and postfix, which is used by the workflow management system to loop through all provided parameters. This strict naming convention requires cumbersome and error-prone manual steps if it has to be repeatedly applied by the user, which is the case for NGS applications. Note that this is not a problem for the workflows executed via the customized gateways, because it is not visible to the end-users.

To supply multiple parameters or file paths in an easier way, a workflow component has been developed that takes a list of file paths or parameters in one plain text file as input, and splits this list into separate text files with the naming convention of WS-PGRADE and which contain just one of the file paths or parameters each. These split files are then passed on to the next workflow component, using the data handling solution presented above. The workflow component is implemented in several sequence analysis workflows, being called “generate-values-...” or “generate-inputs-...”. These are components are of the WS-PGRADE “generator” type.

10.4 Applications Description

10.4.1 Freesurfer

10.4.1.1 Nature and Relevance

Freesurfer is an automated tool to segment regions in the brain in MRI scans. It reconstructs the brain's cortical surface from structural MRI data, and it can also overlay functional MRI data onto the reconstructed surface.
It is intensively used in neuroscience research to isolate regions of interest in which measurements are performed to characterize brain diseases. For example, the volume of the hippocampus is used as a marker of brain degeneration, or activation in the amygdala is used to differentiate between groups of persons with some kind of addiction and control subjects, i.e., healthy persons.
Freesurfer requires long computing times (24-36h) and is often executed for many (healthy and diseased) subjects in each study; therefore, its adoption for large neuroscience studies has become impractical. The execution of this application on the Dutch grid using MOTEUR has enabled a large number of new studies already. The goal in ER-Flow was to revise, optimize and publish the MOTEUR implementation and port it to WS-PGRADE. Moreover, integration of the science gateways with the data servers used by the neuroscientists is expected to further streamline the usage of this application.

10.4.1.2 Usage

The Freesurfer workflow is exclusively started from the science gateways, both for MOTEUR and WS-PGRADE implementations. The Freesurfer suite is free for academic use, but it requires a license that is managed for the gateway users.

10.4.1.3 Software Details

The Freesurfer suite includes around 40 image analysis steps, and generates a large number of files for each of these steps. The temporary files are stored into directories, and reused in subsequent steps. The output generated by Freesurfer is a tree of files with strong implicit semantics. Runtime arguments can be specified to select a sub-set of the steps.

It is common to manually inspect the quality of the results, which is based on quality information available in the generated files. When necessary, the user can manually repair some intermediate result and repeat part of the subsequent steps. There are predefined types of re-runs, depending on the data repaired by the user; these re-runs have special names: ReconAll (complete processing), Pial (predefined subset), and Param (subset specified as workflow input). The complete output tree of files previously generated needs to be given as input for a Freesurfer re-run.

The software is released as a package that can be used directly as a legacy application in the workflows. It contains a license that is provided as input to the code. This license has been obtained for the VLEMED VO, but any academic user potentially can obtain one. The VLEMED license file is currently hardcoded in the workflow implementation and is thus available to VLEMED members.

10.4.1.4 Workflow Details

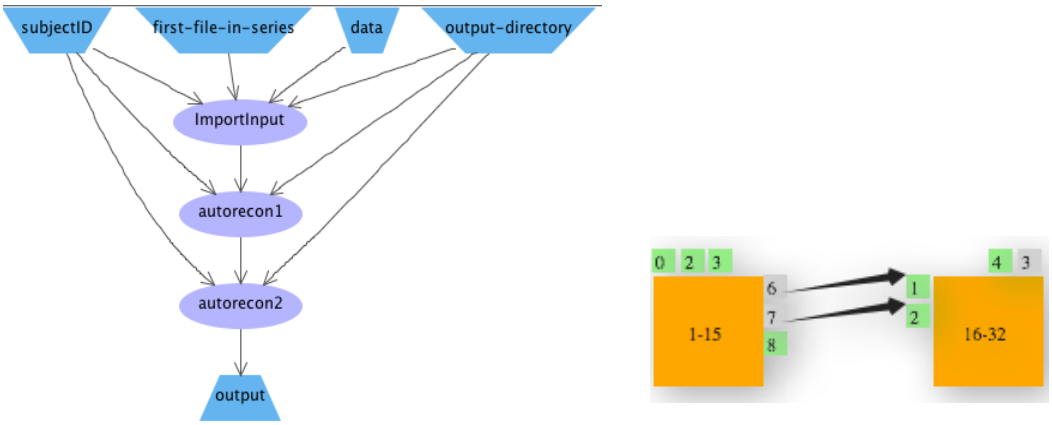

Figure 23, Freesurfer implementations in MOTEUR (left) and WS-PGRADE (right)

The Freesurfer workflow implementations contain two main components:

- Data preparation
- Call to Freesurfer application

Although the Freesurfer application can be executed as one single job, in practice we observed that this was not optimal. Due to job failures, it happened too often that the processing results of many hours of computation were simply lost. The solution we found was to split the job into smaller parts, and force the results to be saved in between, implementing a manual check-pointing mechanism. This approach was used both for MOTEUR and WS-PGRADE implementations, as illustrated in the figure above.

This approach provided an excellent solution for WS-PGRADE: when failures occur, the workflow can be resumed manually, and it will continue from the point where it previously was. This strategy, however, increases the overhead caused by data transfers and job start-up. Therefore, the granularity of the jobs needs to be well tuned for the operational characteristics of our infrastructure. Note that this overhead is less critical in MOTEUR because it uses a pilot job substrate that significantly reduces job start-up overload.

Because the Freesurfer code is so large, it is not part of the workflow bundle description, but downloaded dynamically on the execution node from a grid file. This file bundle also contains the license for Freesurfer execution, which is a text file that has been obtained for the gateway users as a whole.

10.4.1.5 Further Technical Details

Information	URL
SHIWA Repository (MOTEUR)	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=2251
SHIWA Repository (WS-PGRADE)	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4251
Application description	http://www.erflow.eu/documents/388575/774006/ERflow-Application-LS-Freesurfer-v2.pdf
Software Documentation	http://surfer.nmr.mgh.harvard.edu/fswiki
Contact Details	m.jaghouri@amc.uva.nl, vkorkhov@gmail.com

Table 39, Technical details of the Freesurfer application.

10.4.2 DTI Preprocessing

10.4.2.1 Nature and Relevance

Diffusion Tensor Imaging (DTI) has become a popular method to analyse the structure of the human brain, which is composed of grey and white matter tissue. The grey matter is known to facilitate signal processing, and the white matter comprises nerve bundles that connect brain regions to each other. These nerve bundles are macroscopic structures composed of microscopic fibers. DTI measures the local diffusion properties of water in the brain, and can detect these fibers due to their highly anisotropic diffusion properties. Water diffusion is high along the fibre orientation and low (restricted) in the perpendicular direction. This allows for reconstructing both the orientation and integrity of white matter, making it possible to study brain diseases affecting the white matter *in vivo*.

DTI data acquisition is performed in 12 to 60 three-dimensional (3D) orientations, resulting in a series of 3D volumes, in addition to an anatomical scan of the brain. These raw data cannot be directly interpreted, but need complex and computationally demanding analysis. Different toolboxes are available today for analysis of the DTI scans (after being pre-processed), for example FSL BedpostX (see section 10.4.3) and DTI Population Registration (see section 10.4.4).

The DTI preprocessing application implements a sequence of steps that prepare raw DTI data to be used in further analysis packages. The acquisition of a DTI dataset commonly takes 10 minutes, so patients might have moved during this time and motion correction is needed. In addition the data may contain artefacts and noise, which also have to be corrected or compensated for. Moreover, these packages normally require input data to be stored into a special file format and directory structure, as well as submitted to some additional preprocessing steps.

10.4.2.2 Usage

This application is available both on the SHIWA repository and the local AMC gateway. The usage of this application at AMC is exclusively from the gateway by the neuroscientists.

10.4.2.3 Software Details

The DTI preprocessing workflow contains a core component that processes the input data, performs all image analysis steps and generates the results in different formats for further processing. The core processing is performed by a tool developed at the AMC in Matlab, and which has been compiled for license-free execution; therefore, the application has dependencies on Matlab runtime libraries. The input data is a DTI scan in one of the commonly accepted medical image formats (e.g., DICOM and NIFTI). The output is an archive (tar) directory containing various files with quality information about the performed steps and the resulting scan. The location of input and output files are given as logical file names (LFN). The preprocessing is generic, and the generated output can be used for several analysis methods, including BedpostX and DTI Population Registration.

10.4.2.4 Workflow Details

The workflow has been implemented for various DCIs and VOs, since it originates from experiments carried out still during the SHIWA project. There are two main variations, which are illustrated in the figures below:

1. a single-component workflow that generates all output files bundled into a single tar file;
2. an extended version of the workflow with an additional component that extracts BedpostX output and sends it to a separate output port. This version of the workflow can be used directly in conjunction with FSL BedpostX workflow, serving it with the generated data in BedpostX format.

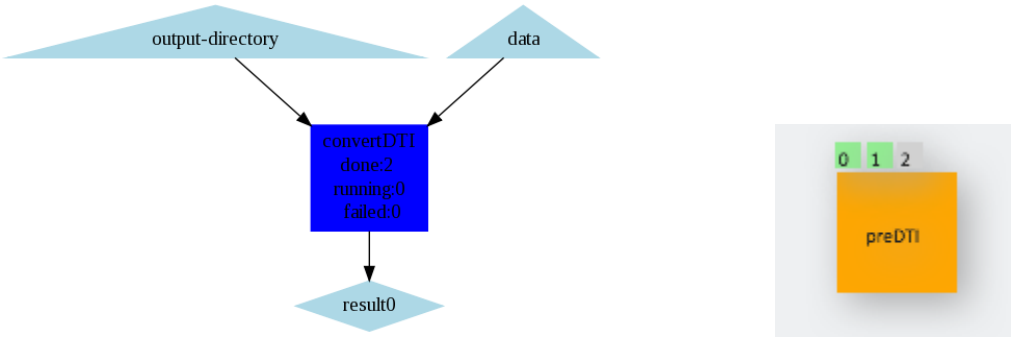

Figure 24,Single-component version of DTI preprocessing workflow, implementations in MOTEUR (left) and WS-PGRADE (right)

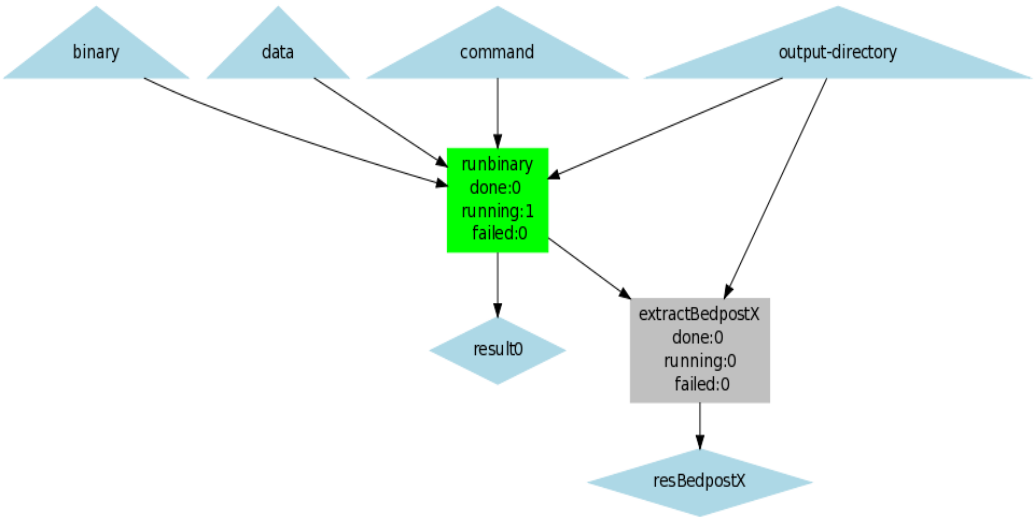

Figure 25, Two-component implementation of DTI preprocessing workflow in MOTEUR

10.4.2.5 Further Technical Details

Information	URL
SHIWA Repository (MOTEUR)	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=1751
SHIWA Repository (WS-PGRADE)	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4250
Software Documentation	http://www.bioinformaticslaboratory.nl/twiki/bin/view/EBioScience/PredtiUserDoc
Application description	http://www.erflow.eu/documents/388575/774006/ERflow-Application-LS-DTIPreProcessing.pdf
Contact Details	m.jaghouri@amc.uva.nl, vkorkhov@gmail.com

Table 40, Technical details of the DTi preprocessing application.

10.4.3 FSL BedpostX

10.4.3.1 Nature and Relevance

DTI data provides information about connectivity of brain white matter (see introduction about DTI in Section 10.4.2.1). The FMRIB's Diffusion Toolbox of the FMRIB Software Library (FSL) developed at Oxford University is one of the toolboxes that are available today for analysis of the DTI scans. In particular, the FSL BedpostX tool is able to detect crossing fibers, which is crucial for correct reconstruction of the white matter structure. For example, this is important for neurosurgery planning.

The method uses a Markov Chain Monte Carlo-based model and is computationally demanding. The processing time of a typical patient's dataset is about 20 hours on a standard desktop computer. When studies are carried out, typically encompassing in the order of 100 subjects, the compute-time sums up to several weeks.

10.4.3.2 Usage

This application is available both on the SHIWA repository and the local AMC gateway. The usage of this application at AMC is exclusively from the gateways by the neuroscientists.

10.4.3.3 Software Details

Two levels of parallelism are possible to speed up execution time. The first is at the level of a single dataset, where each 2D slice of a 3D volume can be processed independently with a gain of a few minutes. Parallelism can additionally be realized for large studies at subject level, by processing the different subjects simultaneously.

The raw DTI data needs to be prepared with the DTI preprocessing application (see section 10.4.2) before it can be used with the FSL BedpostX application. The application is based on FSL library. The computing nodes used for execution of the application must have FSL version 4.1 installed. This package has been installed and is maintained on the nodes of the administrator of the Dutch grid infrastructure.

10.4.3.4 Workflow Details

The workflow executes FSL BedpostX for a single dataset and speeds up the total runtime by splitting the input 3D-volume to 2D-slices and processing them in parallel. It consists of three components: data splitting with the FSL method *fslsplit*, the *fsl-bedpostx* method itself, and data merging using the FSL method *fslmerge*. The input is a compressed archive (tar, gzip) containing a directory with preprocessed data following BedpostX conventions.

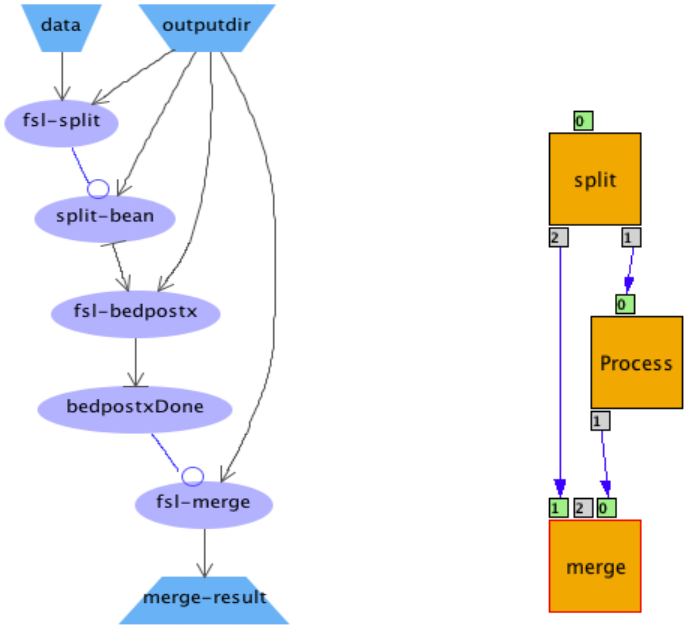

Figure 26, Implementation of FSL BedpostX workflow in MOTEUR (left) and WS-PGRADE (right).

10.4.3.5 Further Technical Details

Information	URL
SHIWA Repository (MOTEUR)	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=1512
SHIWA Repository (WS-PGRADE)	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4858
Application description	http://www.erflow.eu/documents/388575/774006/ERflow-Application-LS-FSLBedpostX.pdf
Software Documentation	http://fsl.fmrib.ox.ac.uk/fsl/fsl4.0/fdt/fdt_bedpostx.html
Contact Details	m.jaghouri@amc.uva.nl, vkorkhov@gmail.com

Table 41, Technical details of the FSL BedpostX application.

10.4.4 DTI Population Registration

10.4.4.1 Nature and Relevance

In the study of the brain, it is necessary to use large collections of scans to obtain measurements that are representative for a particular population. A population contains a group of patients (e.g., displaying depression, schizophrenia, Amyotrophic Lateral Sclerosis, etc.) and control subjects. The scans in this case belong to different people; therefore, they need to be aligned to each other before the measurements (markers) can be compared.

The DTI Population Registration application determines spatial correspondence in DTI datasets of all subjects in a certain population. Typically 50 subjects are included, but larger populations are expected with the growth of study sizes. In this application the "mean" of all subjects serves as template, towards which all subjects are aligned (registered).

10.4.4.2 Usage

This workflow is available on the MOTEUR gateway of AMC. It is also available on the SHIWA repository and therefore accessible from SSP.

10.4.4.3 Software Details

The application is based on the DTITK toolkit. For more information, please refer to the website (http://www.nitrc.org/projects/dtitk).

10.4.4.4 Workflow Details

In this workflow, the mean of all subjects serves as template, towards all subjects are registered. This mean template is iteratively updated until convergence (6 times is optimal). The workflow is illustrated in Figure 27

10.4.4.5 Further Technical Details

Information	URL
SHIWA Repository	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4601
Application description	http://www.erflow.eu/documents/388575/774006/ERflow-Application-LS-DTIPopulationRegistration.pdf
Software Documentation	http://www.bioinformaticslaboratory.nl/twiki/bin/view/EBioScience/DTIPopulationRegistration
Contact Details	m.jaghouri@amc.uva.nl, vkorkhov@gmail.com

Table 42, Technical details of the DTI Population registration application.

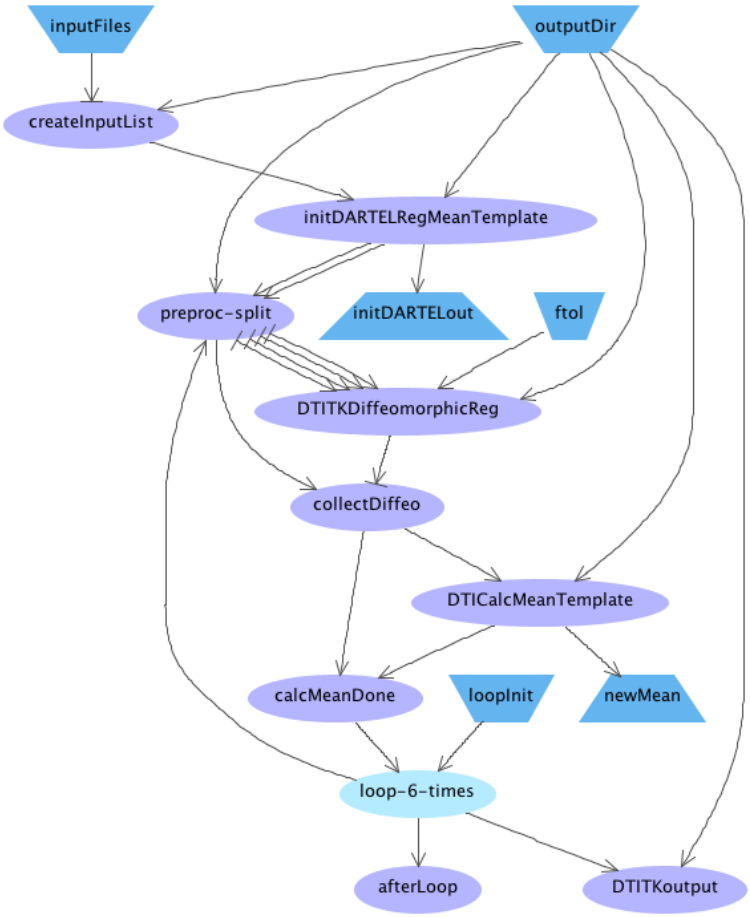

Figure 27, Implementation of DTI Population Registration workflow in MOTEUR

10.4.5 Double Cross Validation

10.4.5.1 Nature and Relevance

This application is an implementation of the repeated double cross validation algorithm, intended for (a) optimizing the complexity of regression models, and (b) for a realistic estimation of prediction errors when the model is applied to new cases (within the population of the data used). In short, in this approach, given a list of x and y variables, we want to find a linear model describing this data as: $y = b_0 + b_1x_1 + ... + b_nx_n$

The main available implementation uses *logistic regression* in order to find such a model from a given set of x and y variables. However, other methods such as support vector machines, PLSDA, and Nonlinear Canonical Correlation Analysis can also be considered.

The purpose of creating such a model is to be able to predict the y values. Typically, y values could refer to diseased or healthy people. In such a case, the goal is to judge whether a person has a particular disease or not.

Building a model from the available data can result in bias towards the data set at hand, i.e., it cannot predict properly, because always the number of subjects considered in an experiment is very little compared to the total population of humans. In order to avoid this bias, a cross validation step works as follows: part of the data is set aside as validation set, and is used to test how accurately the model, generated from the rest of the data, can predict the validation set. By varying the validation set among the total amount of available data, one can obtain different models with different accuracies, and usually the best model is selected. In *double* cross validation, an extra testing phase happens in order to score the best models resulting from the cross validation.

Finally, in order to be able to compare the results, the double cross validation is repeated after applying different permutations on the input data. Statistically, after applying a random permutation, the data should become non-sense and therefore the generated models should not be able to predict properly. Such repetitions can happen for hundreds to thousands of permutations.

10.4.5.2 Usage

This workflow is available on the MOTEUR gateway of AMC. It can also be executed from AMC SCI-BUS portal, but currently no custom user interface has been developed for it. The MOTEUR version is not exported to SHIWA repository because of the lack of automatic export from MOTEUR desktop. There is also a WS-PGRADE version available which is exported to the SHIWA repository and therefore accessible also from SSP.

10.4.5.3 Software Details

The code to calculate the penalized logistic regression models is developed in Matlab and compiled to execute license-free. It therefore requires Matlab runtime to be available on the execution environment. This requirement is met on the Dutch grid infrastructure, although this may change over time and needs to be maintained. Other methods are implemented in R, and the same situation applies to the R runtime environment. The required non-standard R packages used in each method are bundled together with the executable.

10.4.5.4 Workflow Details

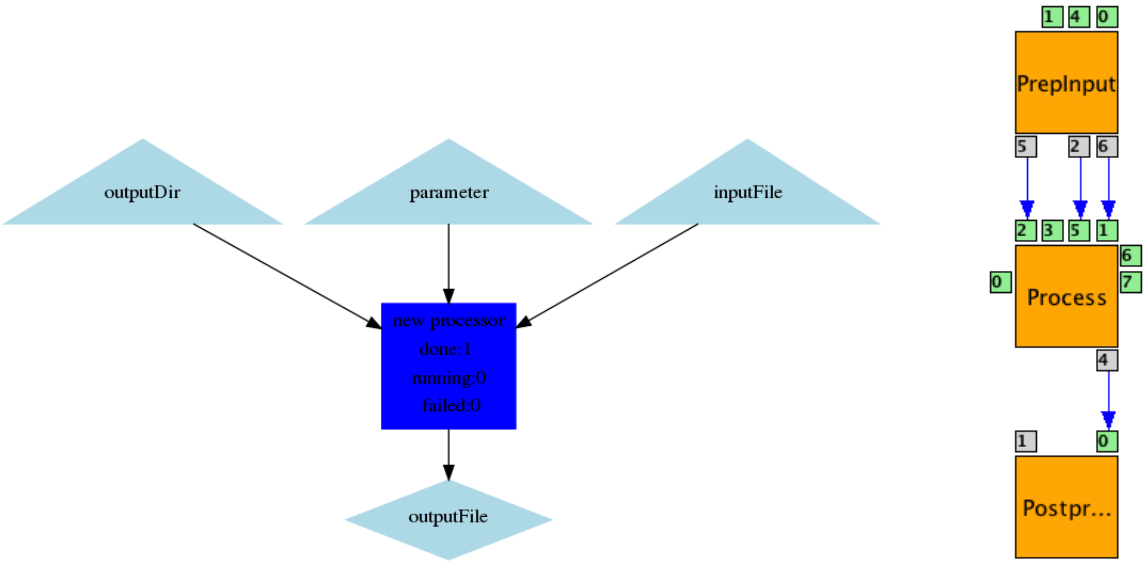

Figure 28, Implementation of Double Cross Validation in MOTEUR (left) and WS-PGRADE (right)

This workflow has the possibility of parallelism at three levels: the two cross validation steps and another level to apply permutations. If all three levels of parallelism are implemented, it will result in thousands of very small tasks. It is well known that grid tasks should not be too small; otherwise the queuing time might out weight the benefits of parallelism. Therefore, in the current WS-PGRADE implementation, we exploit only two levels of parallelism, via parameter sweep (for different permutations) and split-merge (known as generator-collector ports in WS-PGRADE), respectively. In the MOTEUR version, however, only parameter sweep (for different permutations) is used to achieve parallelism.

The WS-PGRADE implementation has an additional feature: it can generate a set of permutations on-the-fly. For the sake of reproducibility of results, it also has the possibility of accepting a list of previously generated permutations as input. The MOTEUR implementation only supports the latter possibility.

10.4.5.5 Further Technical Details

Information	URL
SHIWA Repository (WS-PGRADE)	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4805
Application description	http://www.erflow.eu/documents/388575/774006/ERflow-Application-LS-DCV-v2.pdf
Software Documentation	http://www.bioinformaticslaboratory.nl/twiki/bin/view/EBioScience/LRDCUserDoc
Contact Details	m.jaghouri@amc.uva.nl

Table 43, Technical details of the Double cross validation application.

10.4.6 AutoDock Vina

10.4.6.1 Nature and Relevance

Autodock Vina is a toolbox to perform virtual screening experiments. It finds the preferred orientation of one molecule with respect to another considering a large set of binding affinities. Virtual screens of large databases are used as a starting point to identify small molecules that can interact with proteins and modulate their activity. Potential targets are further evaluated using conventional biochemical assays.
Virtual screening is the first step in modern drug design, however large computing times are necessary to test all possible combinations. At the AMC, as in many other healthcare organizations, there is large interest in investigating new drugs that are more targeted to the disease or to the patient (personalized medicine).

10.4.6.2 Usage

This workflow will be executed from a customized interface to be added to the AMC SCI-BUS science gateway, however the interface implementation has not been completed yet. The workflow is currently ready for usage from the standard interface for workflow developers in the SHIWA Portal and in the AMC SCI-BUS portal.

10.4.6.3 Software Details

AutoDock Vina is an open-source program for doing molecular docking, released under a very permissive Apache license, with few restrictions on commercial or non-commercial use, or on the derivative works. AutoDock Vina achieves an approximately two orders of magnitude speed-up compared with the molecular docking software previously developed AutoDock 4, while also significantly improving the accuracy of the binding mode predictions, judging by our tests on the training set used in AutoDock 4 development. Further speed-up is achieved from parallelism, by using multithreading on multicore machines. AutoDock Vina automatically calculates the grid maps and clusters the results in a way transparent to the user. For its input and output, Vina uses the same PDBQT molecular structure file format used by AutoDock. PDBQT files can be generated (interactively or in batch mode) and viewed using MGLTools. Other files, such as the AutoDock and AutoGrid parameter files (GPF, DPF) and grid map files are not needed.

10.4.6.4 Workflow Details

Figure 29, Implementation of Autodock Vina workflow in WS-PGRADE

This workflow is originated from the desktop grid workflow developed for the University of Westminster AutoDock portal. The original workflow has been published on the SHIWA repository by the authors, and downloaded at the AMC with the intention of reuse. The same application has been also ported by the MoSGrid community for their DCI, which might suggest a joint effort. In practice, however, the details of the DCIs make the implementations very different from each other, and limit chances for exchange of code. The same ideas present in the original workflow developed at UoW could of course be reused.

The general idea of this workflow is that there is a "receptor" that is "docked" against a "database" of "ligands". This is a 1-N data parallelism workflow.
The first component, the "generator", takes the input database and splits it into N pieces.
The second component is the actual AutoDock Vina component. This comprises of a wrapper shell script that sets up some stuff and the actual AutoDock Vina code, which is a statically compiled Linux binary that proved to run everywhere so far.
The last step in the workflow is a “collector” that takes all the outputs of the individual jobs, filters the “M” best results, and creates that as output.

10.4.6.5 Further Technical Details

Information	URL
SHIWA Repository (WS-PGRADE)	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4256
Original DG workflow?	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=2956
Application description	http://www.erflow.eu/documents/388575/774006/ERflow-Application-LS-Vina-v2.pdf
Software Documentation	http://vina.scripps.edu/tutorial.html
Contact Details	vkorkhov@gmail.com

Table 44, Technical details of the Autodock Vina application.

10.4.7 Sequence Alignment (BWA)

10.4.7.1 Nature and Relevance

A sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns.

A large number of tools exist for sequence alignment. This application uses BWA, which is based on the Burrows-Wheeler Transform. This application aligns sequence fragments in fast format to a reference database.

10.4.7.2 Software Details

BWA keeps the reference database in memory, which takes around 4.5 GB for the human genome. The run time scales approximately linear with the amount of input sequences.

10.4.7.3 Workflow Details

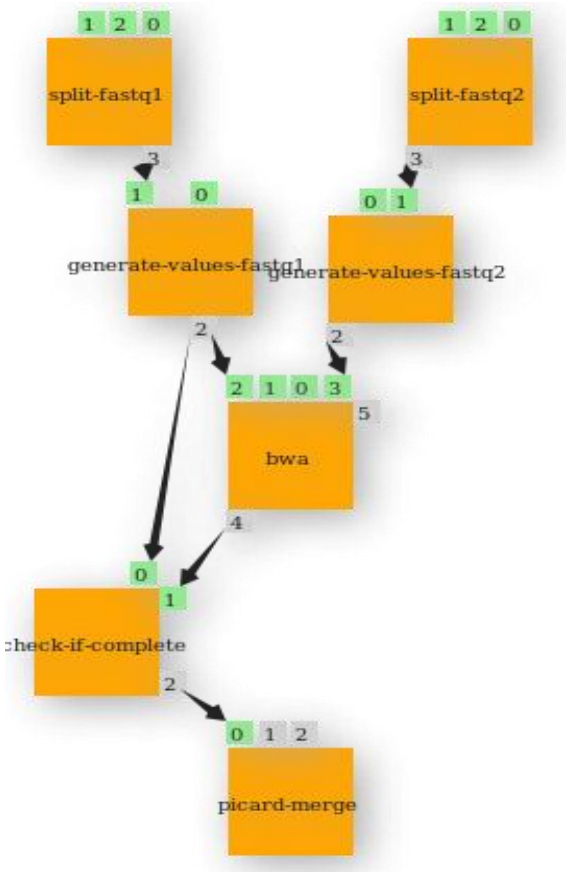

Figure 30, Implementation of BWA workflow for sequence alignment in WS-PGRADE

The sequence reads have to be stored on the computing platform (grid, pbs or local). The inputs supplied to this workflow are plain text files containing the path to the input files, as explained in section 10.3.1.1. The input files are split in smaller chunks, after which they are passed on to the “bwa” component. When all sequences were processed by BWA in parallel,

a component checks if all results (alignment files in bam format) were produced. After this step, they are merged into one alignment file in bam format.

The workflow processes one sequence experiment at the time using data parallelism to speed up the work. If this workflow is used as a sub-workflow it could process multiple sequence experiments at once.

10.4.7.4 Further Technical Details

Further details about BWA is available on the website (http://bio-bwa.sourceforge.net/), and scientific details can be read in the paper by Li H. and Durbin R. (2009), *Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60 [PMID:19451168]*. We have made use of Picard tools, which is available at http://picard.sourceforge.net.

The reference genome (fasta) and bwa index files are archived using tar zcvf. The reference genome used in our experiments is available on:
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz

A BWA index can be build using these instructions:
http://bio-bwa.sourceforge.net/bwa.shtml

Information	URL
SHIWA Repository (WS-PGRADE)	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=1513
Application description	http://www.erflow.eu/documents/388575/774006/ERflow-Application-LS-SequenceAlignment-v1.pdf
Software Documentation	http://bio-bwa.sourceforge.net/
Contact Details	b.d.vanschaik@amc.uva.nl

Table 45, Technical details of the Sequence Alignment application.

10.4.8 SNP Calling

10.4.8.1 Nature and Relevance

One of the goals in next generation sequencing (NGS) is to find variants in DNA data of individuals compared to a reference sequence (genome). This is referred to as “SNP calling”, which implies finding "SNPs" - Single nucleotide polymorphisms, or variants - in NGS data. The extraction of SNPs from the raw genetic sequences involves many processing steps and the application of a diverse set of tools
The raw DNA sequence data is aligned (mapped) to the reference genome (see the “Sequence alignment” application). After the alignment, variants can be determined with respect to the reference.

This application calls variants from (human) genome re-sequencing data with the programs 'samtools' and 'varscan'. The samtools program calls raw variants from a dataset, after which the varscan program determines which SNP calls have more evidence to be true positive calls.

10.4.8.2 Software Details

Both 'samtools' and 'varscan' process alignment files in a linear fashion. An improvement on this workflow could be to split the alignment file per chromosome and process each file in parallel.

10.4.8.3 Workflow Details

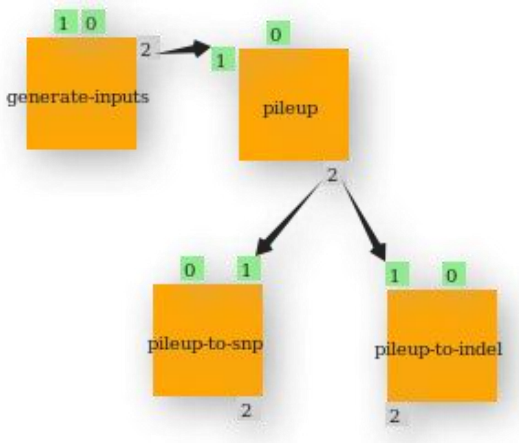

Figure 31, Implementation of samtools and varscan for SNP calling in WS-PGRADE

The alignment files in bam format have to be stored on the computing platform (grid/g-lite or pbs). The inputs supplied to this workflow are plain text files containing the path to the input files. Multiple SNP files can be processed in one workflow run.

10.4.8.4 Further Technical Details

The tools used in this workflow are part of various software packages.

1) Samtools - http://samtools.sourceforge.net/
Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

2) Varscan - http://varscan.sourceforge.net/
Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, & Ding L (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics (Oxford, England), 25* (17), 2283-5 PMID:19542151
Koboldt, D., Zhang, Q., Larson, D., Shen, D., McLellan, M., Lin, L., Miller, C., Mardis, E., Ding, L., & Wilson, R. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing *Genome Research* DOI: 10.1101/gr.129684.111

Information	URL
SHIWA Repository (WS-PGRADE)	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4255
Application Description	http://www.erflow.eu/documents/388575/774006/ERflow-Application-LS-SNPCalling-v1.pdf
Software Documentation	http://samtools.sourceforge.net/, http://varscan.sourceforge.net/
Contact Details	b.d.vanschaik@amc.uva.nl

Table 46, Technical details of the SNP calling application.

10.4.9 SNP annotation

10.4.9.1 Nature and Relevance

To retrieve more information from lists of single nucleotide polymorphisms (“SNPs” or “variants”) it is useful to determine whether the variation is also present in other individuals, or known to be associated with disease. This is done by annotating them with information from public databases, such as gene name and location. The Annovar program is one of the programs to integrate such annotations with SNPs.

The SNP annotation application can annotate variants that were called by the 'varscan' program. It first performs a data conversion step to make the files suitable for 'annovar'. Annotations are added to SNPs using public databases.

10.4.9.2 Software Details

The Annovar program is one of the programs to integrate such annotations with SNPs. Annovar processes SNP input files linearly.

10.4.9.3 Workflow Details

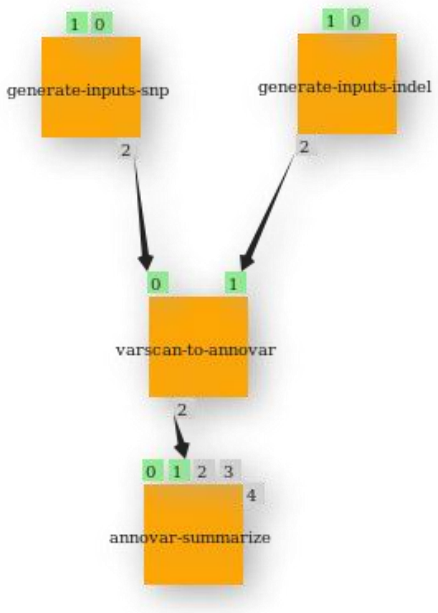

Figure 32, Implementation of annovar for SNP annotation in WS-PGRADE

SNP annotation files are stored on the computing infrastructure and paths to these files are stored in plain text files and supplied to the workflow. The application can annotate multiple SNP files in one go. The 'annovar-summarize' component additionally needs public datasets for the annotation, which is also stored on the computing infrastructure and with the path provided in a text file.

10.4.9.4 Further Technical Details

Annovar - http://www.openbioinformatics.org/annovar/
Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data **Nucleic Acids Research**, 38:e164, 2010

Annotation databases were downloaded using these instructions:
http://www.openbioinformatics.org/annovar/annovar_db.html
All information is downloaded in a directory called 'humandb'. This directory has been compressed with tar zcvf and stored on grid/cluster storage.

Finally, the output of the 'SNP calling' application can be used as input for this application.

Information	URL
SHIWA Repository (WS-PGRADE)	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4254
Application description	http://www.erflow.eu/documents/388575/774006/ERflow-Application-LS-SNPAnnotation-v1.pdf
Software Documentation	http://www.openbioinformatics.org/annovar/
Contact Details	b.d.vanschaik@amc.uva.nl

Table 47, Technical details of the SNP annotation application.

10.4.10 Sample Random Alignments From Alignment Files (DownSample)

10.4.10.1 Nature and Relevance

Random sampling alignments from a sequence alignment file can be used to simulate low coverage sequence experiments. This information can for example be used to analyse whether a sample has been sufficiently covered by sequence reads. In the 'DownsampleSam' program of the Picard toolkit one can indicate the fraction that needs to be sampled from an alignment file (in bam format). For each alignment in the bam file the program uses this fraction to indicate the 'probability' that this alignment should be added to a new alignment file. This method can be repeated on the same alignment file using the same or different fractions, depending on the research question. Being able to perform a parameter sweep on the data significantly speeds up the analysis.

10.4.10.2 Workflow Details

One or more alignment files in bam format are stored on the computing resource (grid, cluster). The paths to these files are provided in a text document. The desired fraction(s) can be provided in a text file via the 'probability' parameter.

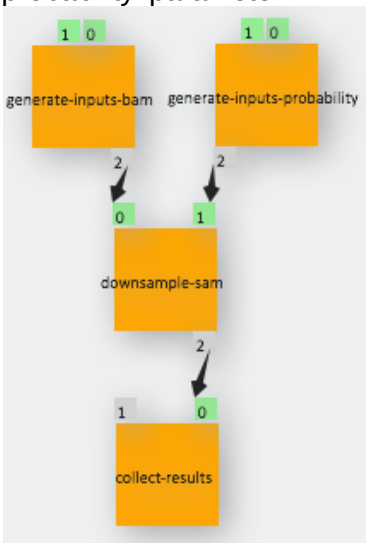

Figure 33, Implementation of DownSample Workflow in WS-PGRADE

10.4.10.3 Further Technical Details

This software makes use of the Picard toolkit.

Information	URL
SHIWA Repository (WS-PGRADE)	http://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/public/edit-application.xhtml?appid=4253
Application description	http://www.erflow.eu/documents/388575/774006/ERflow-Application-LS-LowSampling-v1.pdf
Software Documentation	http://picard.sourceforge.net/
Contact Details	b.d.vanschaik@amc.uva.nl

Table 48, Technical details of the SNP annotation application.

10.4.11 Sequence Assembly

10.4.11.1 Nature and Relevance

Genomes for which a reference genome is not available can be reconstructed using a method called sequence assembly. The overlap between sequence fragments from the studied species is determined, allowing for mismatches between the fragments. A consensus genome is built and a report is made that contains the quality of the assembly and information about the coverage of the consensus genome. The application can assemble sequences in the Roche SFF format.

10.4.11.2 Software Details

Sequence assembly can be a memory intensive process, but this depends on the total size of the studied genome. This application can be used to assemble relatively small genomes, such as viral and bacterial genomes. Depending on the available memory on the computing infrastructure larger genomes can be assembled with this application.

The Roche 'runAssembly' program is implemented in this application, which is also known as 'Newbler'.

10.4.11.3 Workflow Details

Sequences are stored on the computing resources in SFF format. The path(s) are stored in a text file and provided as input to the application. Multiple sff files can be processed in parallel.

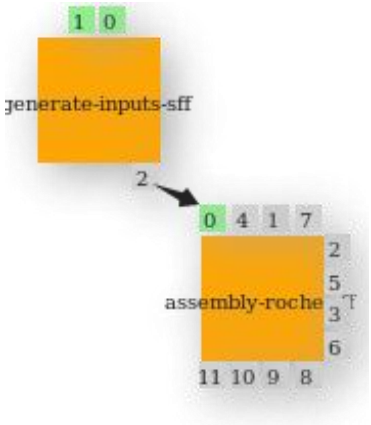

Figure 34, Implementation of Newbler for sequence assembly in WS-PGRADE

10.4.11.4 Further Technical Details

The Assembly software of Roche can be obtained via:
http://454.com/products/analysis-software/index.asp
Note that this software package is not open source, but can be used for free for research purposes. Check the license on the website of the vendor before use.

Information	URL
SHIWA Repository (WS-PGRADE)	Not exported yet, due to a bug
Application description	http://www.erflow.eu/documents/388575/774006/ERflow-Application-LS-SequenceAssembly-v1.pdf
Software Documentation	http://454.com/products/analysis-software/index.asp
Contact Details	b.d.vanschaik@amc.uva.nl

Table 49, Technical details of the SNP annotation application.

10.5 Porting Experience

Three persons were involved in the porting of applications as workflows. Only one had previous experience with the SHIWA platform, and none had experience with WS-PGRADE. In some cases the starting point was an existing MOTEUR workflow, in others the workflow was started from scratch. All new workflows were developed for WS-PGRADE.

The following difficulties were faced during development of the WS-PGRADE workflows. Firstly, the AMC SCIBUS portal became fully operational only in 2013, due to some installation problems. Secondly, we faced various problems related to gLite resources, some of them due to bugs in WS-PGRADE, others due to differences between MOTEUR and WS-PGRADE when using these resources. Additionally, it is very difficult to debug workflows, so the development and porting process was slower than expected.

When the workflows became ready, the next step was to port them to the SHIWA platform. This involved two phases: uploading of workflows to the SHIWA repository, and then running these workflows from the various execution environments for the usage scenarios. Various difficulties were faced in this phase as well, which were solved in the various upgrades both in WS-PGRADE and the SHIWA repository and portal in the past months. Some problems remain, but we trust they will be solved by WP3 in the second project year.

11 Conclusions

In this deliverable the porting of a total of 33 applications have been described. They represent the efforts of four communities to port relevant scientific code 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 science gateways for these communities.

In the process of collecting detailed information for this deliverable, we detected a large heterogeneity among the communities regarding workflow development and documentation styles. For example, MoSGrid chose to use abstract diagrams as visual representations for their workflows, both in this deliverable and in the SHIWA repository. The other three communities used screenshots of the workflow graphs taken directly from the user interface of the workflow management system. Whereas the diagrams are easier to interpret by a scientist, the screenshots are possibly more informative for workflow developers that are interested in reusing the workflow. Although in ER-flow we strive for best practices that can be reused and shared, differences between community cultures should be preserved.

At the end of the first year of application porting activities carried out by the four user communities in WP5 of ER-flow, we can summarize that these activities achieved the expected goals according to scientific and technological relevance, porting activities, and technology awareness.

Concerning *scientific and technological relevance,* we observed that each community has (1) performed an assessment of the current state-of-the art of workflows and applications in their own domain, and (2) has produced a strategy for the creation, adaptation and porting activities that lead to the successful porting of a large number of applications to the SHIWA 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 other platforms (Taverna, MOTEUR, etc.). Each community has selected a set of workflows that were meaningful to the scientific community and that also represented a significant test bed for the technologies adopted in the ER-FLOW project.

Concerning *porting activities*, each community has published a large number of applications into the SHIWA Repository and executed these through the SHIWA Simulation Platform or from the customized science gateways. The porting activities have highlighted several technical aspects that had to be faced for a successful completion. These technical activities will allow smoother porting activities for the second project year.

Finally, a significant outcome of WP5 activities (together with WP2) has been an *increased awareness* note only about *the SHIWA technology*, but also about the *usage of workflow management technologies* to port applications to Distributed Computing Infrastructures. This increased awareness has been particularly noticeable for the communities that were relatively new to workflows, such as Astrophysics and Heliophysics. The close cooperation between technical people and the scientific communities has fostered the creation of cross-communities with two main beneficial outcomes. On the one hand, researchers involved in the technical development of the SHIWA platform have developed a better understanding of the interests of the scientific community and now have a better understanding of their practices and needs. On the other hand, the scientific communities have now a better understanding of the possibilities offered by workflows technologies in general and SHIWA in particular. This mutual understanding will be beneficial for activities in the second project year.