IS-ENES – WP3

D 3.5 – The suite of reference workflows

Abstract

Grant Agreement number	**228203**	**Proposal Number:**	**FP7-INFRA-2008-1.1.2.21**
Project Acronym:	**IS-ENES**		
Project Coordinator:	**Dr Sylvie JOUSSAUME**		

Document Title:	The suite of reference workflows			**Deliverable:**	D3.5
Document Id N°:	D3.5	**Version:**	1.0	**Date:**	12.03.2012
Status:	Final				

Filename:	ISENES_D3.5_final
Authors:	Heinrich Widmann

Project Classification:	Public, Confidential

Approval Status		
Document Manager	**Verification Authority**	**Project Approval**

REVISION TABLE

Version	Date	Comments	Authors, contributors, reviewers
1.0	Mar, 12 2012	Initial Version	Initial Version
2.0	February 2013	Revised version; reviewers: Reinhard Budich & Marie-Alice Foujols	Revised version; reviewers: Reinhard Budich & Marie-Alice Foujols

Table of Contents

1 Introduction

Earth System Modelling (ESM) is becoming very demanding not only due to the growing complexity of earth system models, but also due to the increasing diversity of existing resources such as HPC platforms, grid services or specific data processing tools. Therefor the task to port user- and site-specific ESM applications effectively to the appropriate resources becomes a great challenge.

The aim within the sub task “ESM workflows” of the project IS-ENES is to help and support scientists in mastering this challenge by keeping the focus on science while using best practices. We first analyse the situation by re-examining ESM activities and use cases at ESM centres. For this purpose a questionnaire has been implemented to gather information about real-world as. Especially we are interested in workflow related items as occurring bottlenecks, restrictions and requirements for IT resources. Furthermore we want to know what the typical workflows in ESM are and how they differ each other among different modeling centres. We present the information gathered from this questionnaire and other surveys in the portal.

The analysis of the survey yielded basic ‘reference workflows’, i.e. generalized workflow templates for basic ESM applications. These workflows are presented to the science community in a proper, comprehensible and understandable way. Instead to present a suite of reference workflow, we focus here on the overarching and comprehensive “CMIP5 workflow”, that contains other ‘reference workflows’ as sub-workflows. Scientists will benefit from this representation not only by getting a structured view of the complex underlying processing pipelines, but also by re-using sub-workflows and by referred to the effective usage of tools and services. The overall intention is to simplify composition and enhancement of user-specific workflows, which will result in an effective execution and optimal performance.

For a scientist without detailed technical knowledge it is difficult to identify the processing step, where the congestion occurs. Even more tedious than to localise the bottleneck is to overcome the restriction and to improve the performance without getting lost in technical details. We try to assist the scientist by suggesting alternative processing options, that have been proven to be more efficient for the task in question. For instance, it is known, that bottlenecks often occur during model I/O. In this case the solution maybe parallelisation of the I/O process or the implementation of a dedicated I/O server.

This document focuses on the implementation of the questionnaire and presentation of the results within the IS-ENES portal. Here lies the emphasis on conceptual and technical design issues and is divided into five sections. In the following section we provide the used terminology and the roadmap of the subtask “The IS-ENES ESM workflows”. The acquirement of information about site specific use cases is subject of the third section. The resulting ‘’reference workflows’’ are presented in section four. On the portal page of ‘’The IS-ENES ESM workflows’’ (https://enes.org/computing/workflows) links to the questionnaire, to the results of the surveys and to the reference workflows are provided, see the screenshot in figure 1.

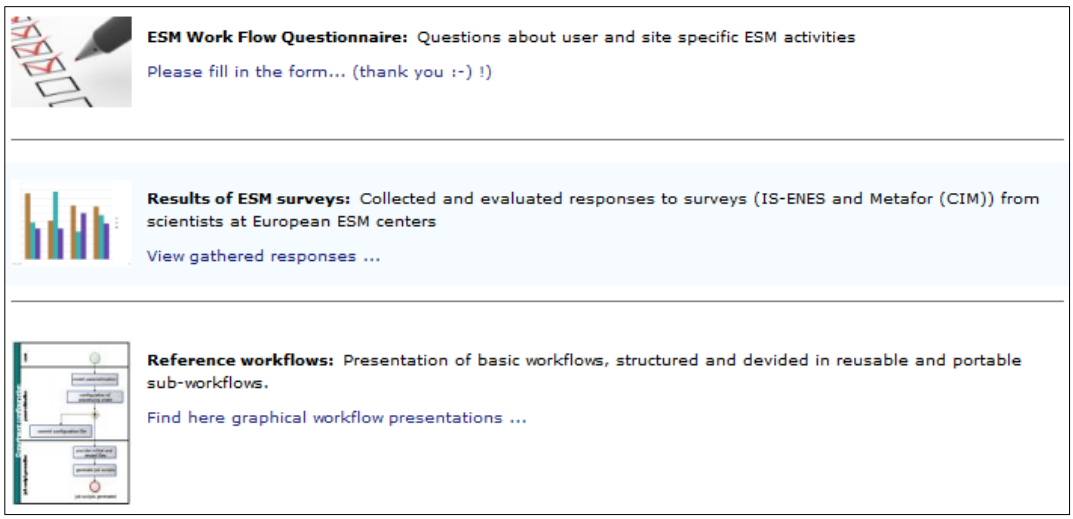

Figure 1 : The main links on the page “The IS-ENES ESM workflows” in the project portal

In section five we discuss workflow management tools and languages, that can be used to build executable workflows. Finally we give a summary and an outlook.

2 Scientific workflows in earth system science

In the subsequent sections we give some definitions in the context of “ESM workflows”, especially we explain, what we mean with the term “scientific workflow”. Based on these definitions and the roadmap described below we developed, structured and presented ESM workflows.

2.1 Nomenclature

One can identify three central issues of ESM research from the overview given on the IS-ENES portal website (see https://enes.org) :

- **Climate Models** or Earth System Models comprise all components of the climate system (atmosphere, ocean, sea-ice and land) and biogeochemical cycles, e.g. the carbon cycle, vegetation and chemical processes. The effective execution of the model on an appropriate HPC or GRID system is hereby the major challenge.
- **Climate Data** can be divided into three groups : input, output and meta data. ESM workflows are extremely data-oriented, i.e. they are triggered by data and data management dominates the processing pipeline from model I/O to storage and visualisation.
- **Climate Computing** resources include hardware as computers or software as services and tools. They are provided by the local computing centre or remote HPC or GRID systems.

In this context of climate research an “ESM workflow” consumes *Climate Computing* resources to perform *Climate Model* simulations, which are forced by and produce *Climate Data*.

Besides this *contextual* definition we refer here mainly to the more *technical* definition of "scientific workflows" to describe and manage complex model and data processing. Scientific workflows are networks of reusable, modular sub-workflows and describe applications on a high-level user view by hiding technical details. This allows scientists to run complex ESM applications and access distributed data sources by sharing reusable sub-workflows.

2.2 Roadmap: From use case to executable workflow

Based on the above given syntax of “sub-workflows”, we define hierarchical workflow levels, to describe the phases of the way from a concrete use case to an executable workflow. Each workflow is split into several sub-workflows on the next level. In figure 2 we sketch the roadmap, that passes through the following phases or levels:

L0. **Use case**: The user describes a specific application or ‘use case’ of his daily work. This can be done e.g. by UML[1], free text or providing commented source code. Of special interest is the description of occurring bottlenecks, used or required processing tools and resources.
L1. **Conceptual view**: The use case is mapped to a conceptual view of the workflow, which gives just a very general and overarching description and splits only in a few main parts.
L2. **Logical view:** Each of the blocks on the *conceptual* level is further divided in more modular sub-workflows. On this logical level typically each sub-workflow can be run isolated on different machines and by different services.
L3. **Processing view**: On this level we describe each workflow in more detail using a processing language, by which we can assign events, tools and conditional processing.

[1] The Unified Modeling Language (UML) is a standardized general-purpose modeling language in the field of object-oriented software engineering.

L4. **Executable level** (not shown in figure 2, for details see section 5.1) : Finally the workflow on the processing level is implemented in a framework or a “workflow engine” to gain an executable, reusable and adaptable workflow.

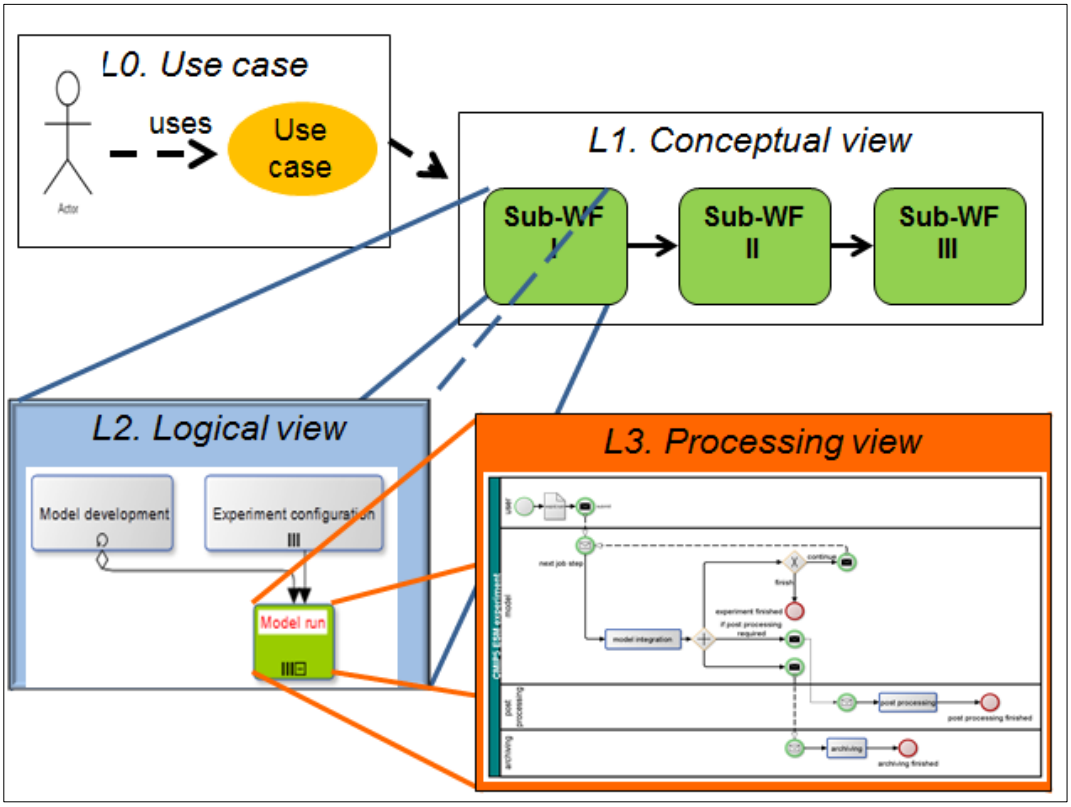

Figure 2 : From use case towards executable workflow

3 Information acquirement

According to the roadmap shown in the previous section the first phase (L0) includes gathering of information about user- and site-specific applications and ‘use cases’ from the ESM community. The methods used to collect this information are described in the following sub sections.

3.1 The workflow questionnaire

We set up a workflow questionnaire, accessible at http://enes.org/computing/workflows/questionnaire. The technical implementation is based on Plone[2] forms. The latter provide a wide range of widgets as multi-select menus, checkbox fields and mailer functionality.

The following screenshots illustrate the layout of the form. The design and set up is orientated to user-friendliness allowing a quick fill-out of the form. We highlighted the benefits of taking part in the survey by incorporating a block “Description and Motivation” at the beginning of the form, as shown in figure 3.

[2] The IS-ENES portal is implemented and developed with the content management and web framework software Zope/Plone, see http://www.plone.org .

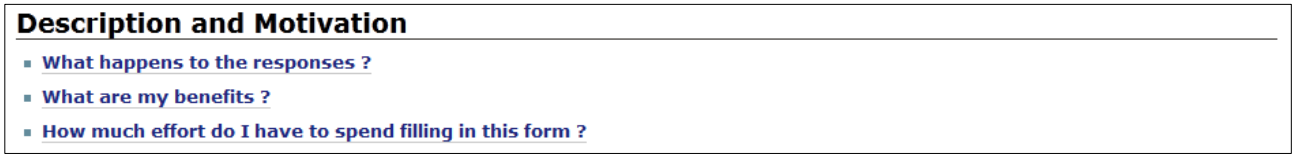

Figure 3 : First block "Description and motivation" of the workflow questionnaire

After asking in the first paragraph "I. Contact, affiliation and job position" for general information (see figure 4),

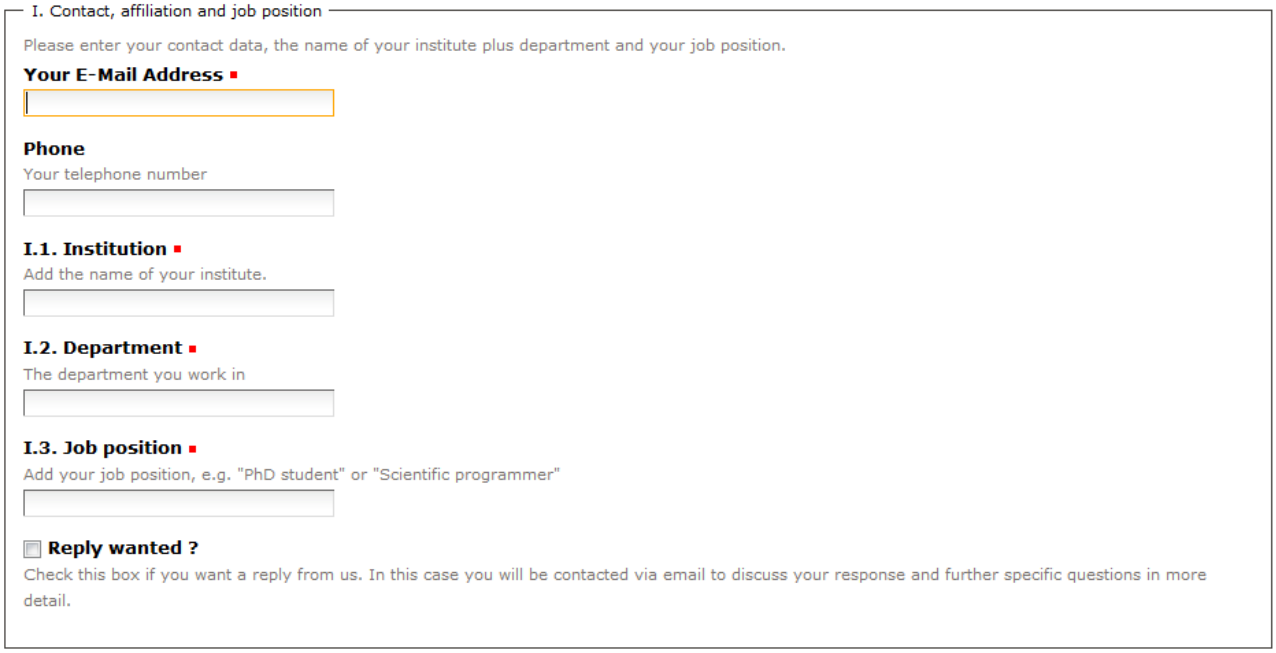

Figure 4 : Paragraph I. Contact, affiliation and job position

the subsequent paragraphs regard content and address specific workflow aspects in more detail.

E.g. in paragraph III (see figure 5) we ask for ESM activities used or scheduled. One can choose an activity from the multi-select menu or describe it in the free-text field.

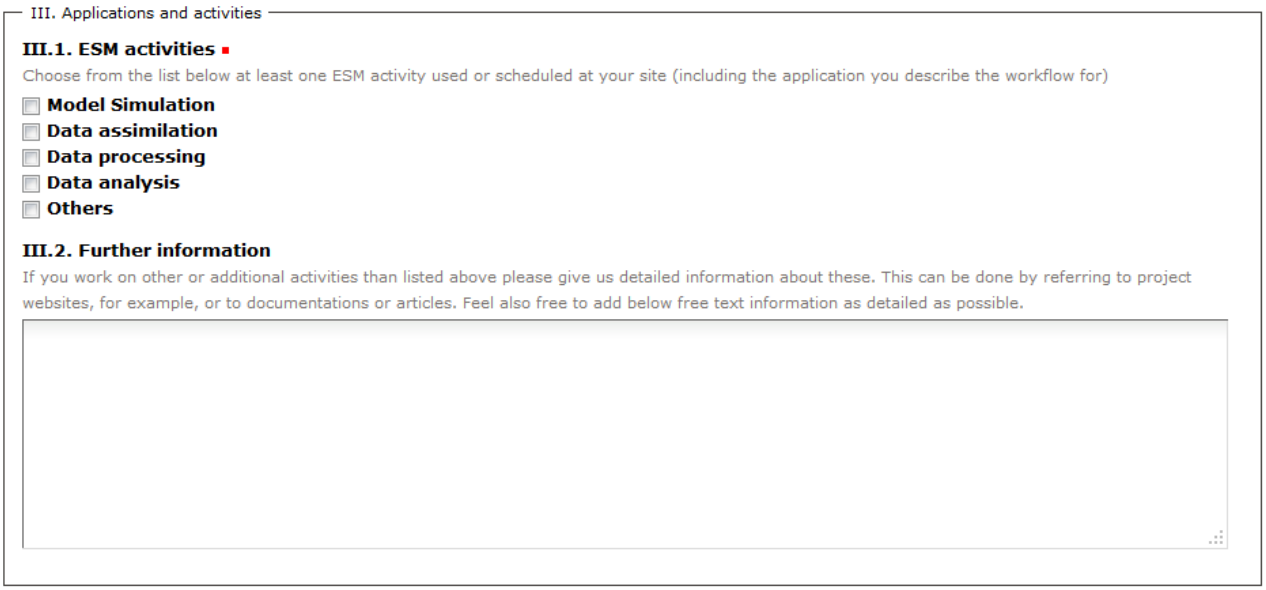

Figure 5 : Pragraph III. Applications and activities

In the subsequent paragraphs IV. to IX. We ask for information such as model configuration, workflow steps, modelling environment, hardware and software resources.

In the table of paragraph “X. Performance” used resources can be assigned to several processing steps, including the activities asked for in paragraph III. As shown in figure 6 for each activity we ask for the consumed computing time, needed disk space and used processing platform.

X. Performance

Give for each workflow step the amount of consumed ressource and on which platform is is performed.

X.1. Consumed ressources and processing platform

The following table needs some explanation : - CPU time should be given as wall clock time in seconds needed to perform one model year - Disk space needed corresponds to the amount of data produced by the process in question within one model year - Platform the workflow step in question is performed on

	cpu time [hh:mm:ss / myr]	disk space needed [mb / myr]	processing platform
Preprocessing			
Model integration			
Assimilation step			
I/O			
Postprocessing			
Data storage			
Data Transfer			

Figure 6 : Questions about consumed resources and processing platform

At the end of the form the opportunity is given to give some final remarks, especially comments and critics of the design and content of the questionnaire are welcome.

When the questionnaire had grown to be quite comprehensive, we invited selected scientists to take part in the survey. Up to now however, there has not sufficient response for a statistical evaluation. Therefor we decided to conduct direct interviews with scientists, to get concrete information of their needs from the scientists.

3.2 Collected results

Additionally we harvested information from other surveys available to us. These are:

- The IS-ENES User Survey: Questions related to performance, coupling and tools in ESM were sent around by email to ESM centres and responses are collected as free text answers in ASCII format.
- The Metafor[3] CMIP5 questionnaire: There exists a strong cooperation between the projects IS-ENES and MetaFor. Especially the meta database CIM (Common Information Metadata), where amongst other things information about CMIP5 models, simulations and platforms are collected, plays a fundamental role in both projects.

Workflow-related responses are collected, grouped by source and presented in the portal at https://enes.org/computing/workflows/results-of-esm-surveys .

[3] MetaFor project site : http://metaforclimate.eu

4 Reference workflows

On the basis of the evaluation and analysis of the information described in section 3 we identify “reference workflows", i.e. templates, which generalize real-world instances. We focus on the outstanding, overarching and comprehensive "CMIP5 workflow” and present it graphically in the portal. To ensure that the provided reference workflow meets the needs of the ENES community, it should include links to information about required resources and make them accessible and useable.

4.1 Graphical presentation within the portal

For visualization and interactive presentation of workflows we chose the graphical editor yEd to create ‘clickable’ SVG figures, that can be integrated to the portal. For each workflow a SVG figure is generated, that may contain sub-workflows. This allows to “unfold” sub-workflows by clicking on the corresponding box in the portal. This is illustrated in more detail in the following subsection by means of the hierarchical sub-workflow structure of the “CMIP5 workflow”.

4.2 The overarching CMIP5 workflow

According to the roadmap levels in subsection 2.2 we describe first the use case “CMIP5” (L0). The complete description of this use case and the CMIP5 project would fill pages. We give here only a summarized and brief description :

- “The whole processing pipeline within a CMIP5 experiment from generation, over evaluation up to storage and publication of data.”

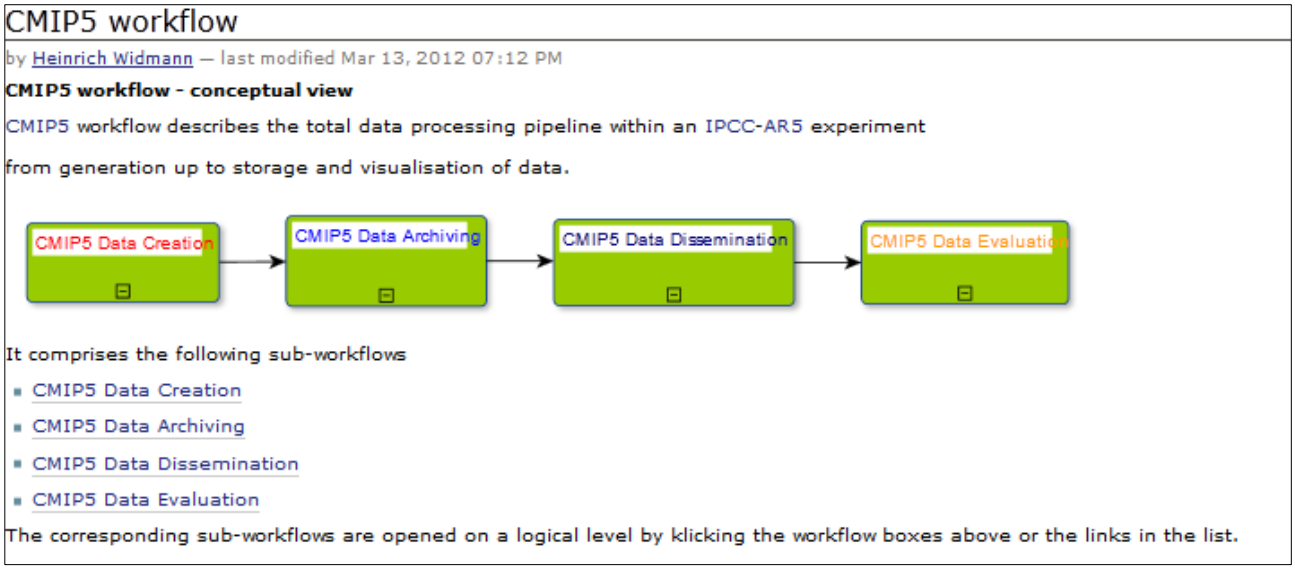

Figure 7 : Presentation of the "CMIP5 workflow" as reference workflow (conceptual level)

The overarching "**CMIP5 workflow**" on the conceptual level (L1) as provided in the IS-ENES portal (see figure 7) comprises the following four sub-workflows on the second logical level (L2) :

- **CMIP5 Data Creation** (L2) describes the process of data generation of CMIP5 experiments, i.e. it corresponds to the process chain of model development (or generation of a model executable), parameterisation of the model run (i.e. experiment configuration) and launching of the model run, that ends with output of raw model data and maybe some standard post processing, see figure 7

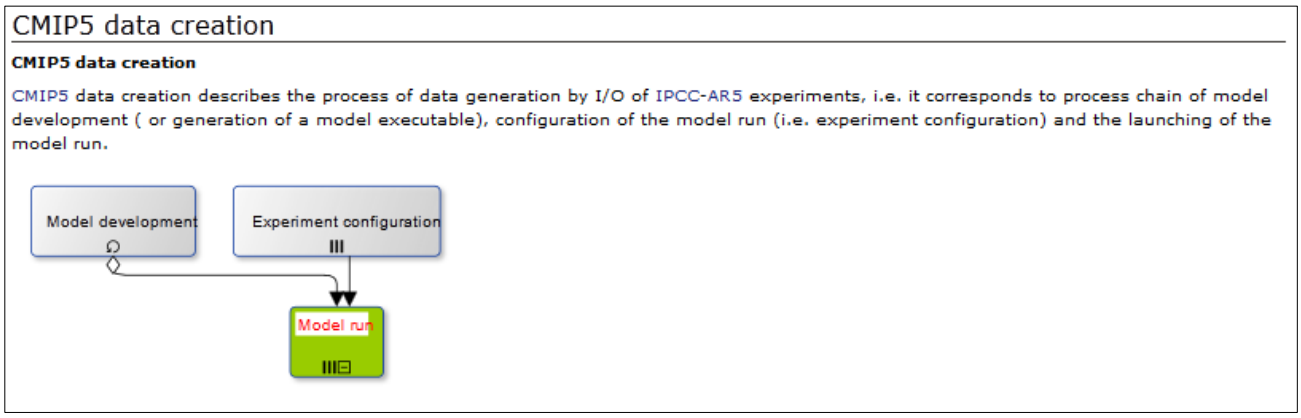

Figure 8 : Presentation of the sub-workflow "CMIP5 Data Creation" on the logical level (L2)

- **CMIP5 Data Archiving** (L2) describes the storage processes of data originating from CMIP5 experiments. For CMIP5 this is essentially the CMOR2 processing.
- **CMIP5 Data Dissemination** (L2) describes the process of quality control, distributed archiving, versioning and publishing CMIP5 data together with managing and providing the accomplished meta data. CMIP5 data are published in the ESG[4] portal.
- **CMIP5 Data Evaluation** (L2) describes the evaluation and processing of CMIP5 data – maybe retrieved from the central ESG archive or available on local discs.

In figure 9 it is shown, how the unfolded workflow "Model run" is presented in the portal on the processing level L3, if one open it by clicking on the corresponding green box (see figure 8).

"Model run" means the execution of a CMIP5 experiment including the launching of the model integration on the computing platform. The accomplished workflow splits again in several sub-tasks :

- User : submission of the job (run script) to a chosen computing platform
- Model : execution of the initial job and submission of subsequent jobs
- Postprocessing : done within the model or direct after saving of model raw output on disc (i.e. processing, which is NOT part of "CMIP5 Data Evaluation")
- Archiving : done immediately after model integration (i.e. archiving, which is NOT part of later on done "CMIP5 Data Archiving"

[4] ESG (Earth System Grid) data nodes, which store some or all parts of CMIP5 data can be accessed at https://enes.org/data/direct-data-access/european-data-nodes/esgf-data-nodes

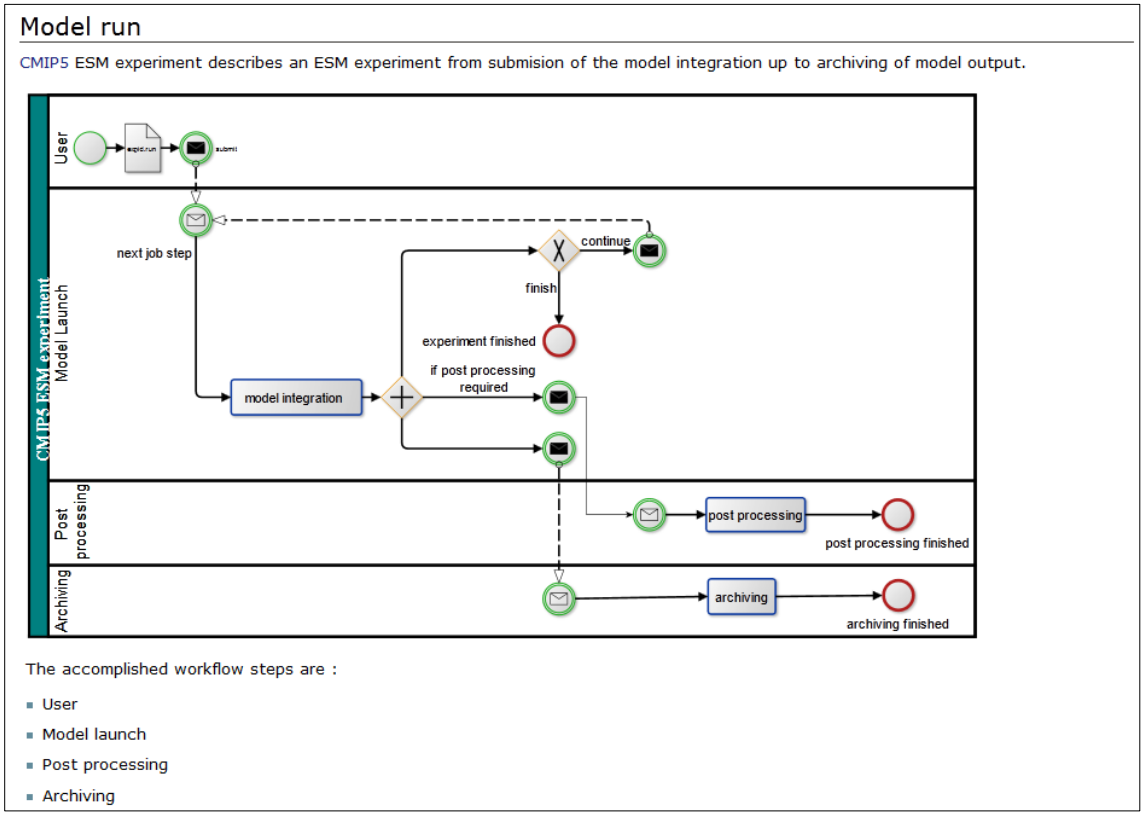

Figure 9 : The sub-workflow "Model run" (processing level)

The last task would be to create an executable workflow (level L5), by implementing the processing workflow within a workflow engine (e.g. within Kepler, as shown below in section 5.1).

4.3 Intended use and best practices

In the context of workflows best practices should simplify the composition and enhancement of user-specific workflows and direct scientists to options, tools and services, which help to overcome bottlenecks and restrictions. We recommend composing and designing first a high-level abstract “scientific workflow” as shown above.

This approach has several advantages:

- reusable sub-workflows are shared within the community and scientist benefit from already known and provided solutions and ‘best practice workflows’,
- each modular sub-workflow can be analysed and conformed to the specific needs on its own,
- for each task scientists will be directed to "HowTos" and maybe to contact persons, who can help and give support for the tool or service in question and
- finally all components can be composed to a network of sub-workflows with transparent and adaptable interfaces.

A typical use case is:

- Launching jobs onto HPC systems : Often there is little support for the usage of the Job Management System. To choose, for instance, the proper queuing directives to get the maximal CPU-power from the computing platform for your model run is a non-trivial task.

For this use case we provide the representation of the reference workflow “Model run” as a sub-workflow of “CMIP5 Data Creation”, as shown in figure 9. We intend to guide scientists from this view on the “processing

level" to an optimised executable workflow (see figures 10 and 11 below), that results in an optimal performance.

This is an example, how scientists are led to use "best practices" and are assisted to map their applications to appropriate HPC resources.

5 Workflow management

In science there is a great interest in tools to manage workflows. We discuss the tool Kepler which is used in ESM research. We also present the workflow language BMPL to structure workflows on a processing level.

5.1 Kepler

We investigated the workflow management system *Kepler*[5] , which is based on the syntax of "scientific workflows". Kepler is a software application for the analysis and modelling of scientific data. It simplifies the effort required to create executable models by using a visual representation of these processes. These representations display the flow of data among discrete analysis and modelling components. Kepler allows scientists to create their own executable scientific workflows by simply dragging and dropping components onto a workflow creation area and connecting the components.

Sub-workflows within Kepler are called "composite actors", because they are further composed of modular workflow components, so called "actors", which may be processes, data or displays. This is a powerful feature to structure nested processing chains and to build comprehensive, stand-alone sub-workflows. These modular components can be exchanged and reused within the community by adapting the input and output ports, which avoids "reinventing the wheel".

For instance the sub-workflow "Model run", as shown in subsection 4.2. on the processing level (L4), splits into further workflow steps. The Kepler presentation in figure 10 shows the order of task execution and the workflow steps that can be performed in parallel (e.g. "Postprocessing" and "Monitoring").

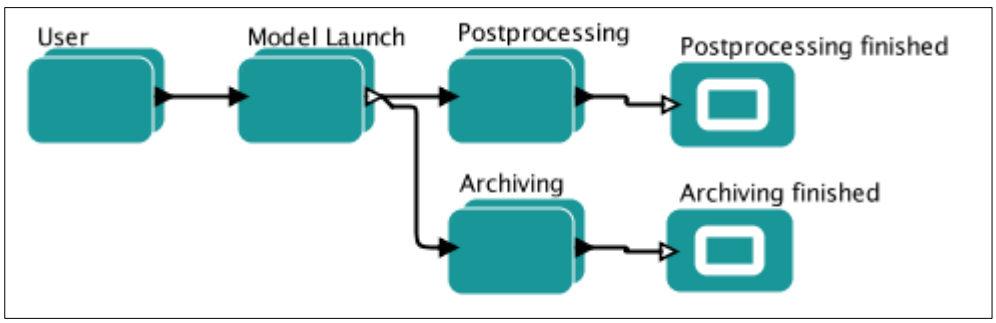

Figure 10: Sub workflow "Model run" as presented within Kepler

The composite actor "Model Launch", showed as folded sub-workflow of "Model run" in figure 10, is shown unfolded in figure 11, where the modular "actors" playing a role during launching the model can be identified.

[5] http://kepler-project.org

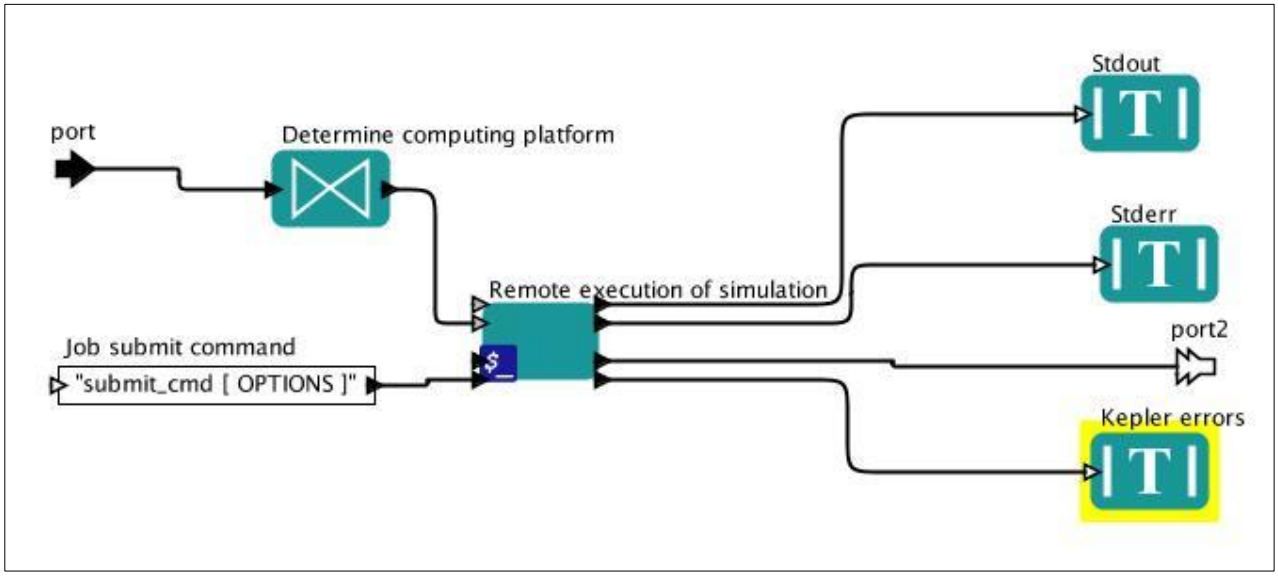

Figure 11 : Opened composite actor "Model launch"

Within Kepler workflows or individual actors are configured by setting parameters. For example, the computing platform can be selected by setting the parameter “Computing platform” to the desired value.

5.2 BPMN

The Business Process Model and Notation (BPMN) is a graphical representation and developed for specifying business processes. This workflow language can be used as well to represent ESM workflows on the processing level (L3), because it provides the necessary elements such as events, activities, gateways and connections. Without going in more detail we refer to the figure 7, which shows, how we use all these elements are used to describe the modular actors within the sub-workflow “Model run” on the processing level.

6 Summary, conclusions and outlook

The implementation of the questionnaire is an iterative process, i.e. from responses and comments from the ESM community we get ideas, which further queries should be added. Because up to now responses are very sparse, we have to convince the community of the benefits of the questionnaire. We will continue to interview scientists directly and point out exemplary applications to help them optimize their executable workflows. Furthermore results from similar surveys are harvested, evaluated and published in the portal.

We showed how tools as Kepler can help to create executable workflows in a structured and understandable way. This report focused on the presentation and analysis of the overarching reference workflow "CMIP5 workflow", because it comprises most sub-workflows occurring in ES research. Further analysis led to first results w.r.t. "best practices".

Based on the results of the present *network activity* we recommend, that it is transformed into *a research activity* of IS-ENES, having the following next steps :

- make reference workflows directly accessible and enable easy creation of executable workflows,
- provide ready, executable workflows, which perform e.g. standardised “data processing” or “data visualisation” in parallel, perhaps on virtual processing nodes or within a GRID environment
- use other features of workflow engines such as Kepler as well, e.g. services to grid-enable applications