EUROPEAN COMMISSION
Community research

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

# IS-ENES2 DELIVERABLE (D -N°: 4.4)

## *Meta-data capture final workshop report*

**File name: {IS-ENES2_D4.4.pdf}**

Author: ***F. Toussaint*** Reviewer: ***R. Budich, M. Carter***

Reporting period: ***01/04/2016 – 31/03/2017***

Date for review: ***15 /02/2017*** Final date of issue: ***23/02/2017***

| Revision table | | | |
|---|---|---|---|
| **Version** | **Date** | **Name** | **Comments** |
| 0.1 | 2016-10-23 | Frank Toussaint | First draft |
| 0.2 | 2017-02-03 | FT | Conclusions enhanced and expanded. |
| 0.3 | 2017-02-23 | RB, MC | Most remarks of the referees inserted. |
| 1.0 | 2017-02-23 | FT | Final corrections |

## Abstract

The deliverable aims at summarising and evaluating the European and global activities addressing developments in metadata capture of both catalogue metadata and adjacent metadata as model descriptions and user annotations. It starts from the analysis of the *Initial workshop on meta-data generation during experiments* (2014-01, Hamburg, see MS 4.2 at https://verc.enes.org/ISENES2/documents/milestones), summarises the results presented at the *Final workshop on meta-data generation during Experiments* (2016-09, Lisbon, see MS 4.6) and draws recommendations from them.

| **Project co-funded by the European Commission's Programme Horizon 2020 (FP7; 2007-2013) under the grant agreement n°312979** | | |
|---|---|---|
| **Dissemination Level** | | |
| **PU** | Public | **X** |
| **PP** | Restricted to other programme participants including the Commission Services | |
| **RE** | Restricted to a group specified by the partners of the **IS-ENES2** project | |
| **CO** | Confidential, only for partners of the **IS-ENES2** project | |

# Table of Contents

# 1 Executive Summary

Background

In the Annex of the EU proposal to the IS-ENES2 project, the work package description of WP4/NA3 aims for four workshops, two on workflow solutions and two on metadata generation during experiments.

The two first workshops took place in Hamburg, Germany, in February and June 2014. They led to two IS-ENES2 milestones and to deliverable D4.2 which summarized the results of the Initial Workshop on workflow solutions (M4.4).

The two final workshops have been merged into one event, the Joint IS-ENES2 Workshop on Workflows and Metadata Generation which took place from 2016-09-27 to 29 in the Hotel Costa de Caparica in Lisbon, Portugal. This document is deliverable D4.4 which provides workshop report of the Metadata (MD) relevant elements of the workshop.

The situation concerning MD creation before the start of IS-ENES is summarised in the description of work for IS-ENES2 WP4/NA3 Task 3: "Significant experience has been gained in CMIP5 and related exercises in providing meta-data to describe ESM experiment sets. A number of sites are recognising the need to build meta-data capture into the heart of the ESM experiment process and to drive data provision exercises; this needs to be supported by both software and processes. This networking activity will promote the sharing of experiences and designs in this emerging area through two workshops organised by DKRZ. The aim will be to encourage investment in software and working processes that will allow more comprehensive meta-data to be collected more efficiently. Further, the development of workflow and diagnostic solutions will be influenced by the metadata requirements." As can be seen, the interaction between workflow and MD was expected and strongly influenced this joint workshop.

The agenda and attendees of the workshop are provided in the Appendix.
Statistics: There were 52 attendees from 17 institutions, of which four institutions were not within the IS-ENES2 participants' list.
The external contributors were:

- Florent Lebeau — Allinea
- V. Balaji — GFDL / NOAA
- Jeffrey Durachta — GFDL / NOAA
- Chandin Wilson — GFDL / NOAA
- Joao Pina — LIP
- Hilary Oliver — NIWA

More detailed information on this event is at
https://verc.enes.org/ISENES2/events/final-is-enes2-workshop-on-workflow-solutions-1

Summary of Results

The combined workshop proved to be very successful. The workflows were evolving with meta-data in mind. This was leading to improved efficiency required to meet the increased complexity of CMIP6. However, this evolution mainly was within single institutes and often did not originate from community cooperation. So there seems still to be potential for synergies.

Further, the relatively high number of participants, probably fostered by the merge of two IS-ENES2 workshops with ESIWACE activities, led to many interesting discussions to which enough time was devoted. Here emphasis was put on sharing of best practices across these two, increasingly related, fields.

In addition to the presentation of many new software (SW) enhancements and developments, hands-on sessions on new software packages (Cylc and AutoSubmit) were offered in parallel to the plenary sessions. This, too, encouraged many participants to attend the workshop.

With the forthcoming challenge in data volumes and complexity from CMIP6 in mind, the developers of workflow and MD tools were given ample evidence of the opportunities to further improve the performance of their systems to meet the MD challenge.

The main outcome of the workshops on metadata generation was that:

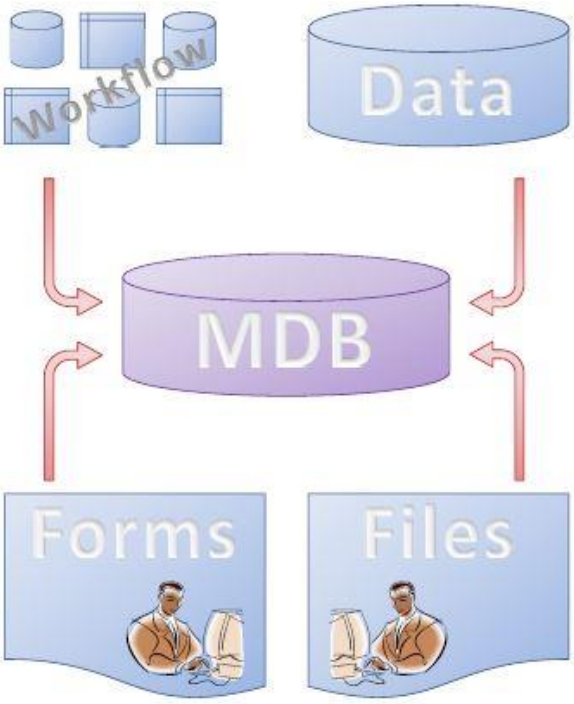
- Despite the improvements made between CMIP5 and CMIP6, the community would benefit in further investment in more robust standardisation for metadata content, structure, and formats.
- Similarly, interfaces standards could be improved.
- For work with metadata and homogenisation of metadata, efforts of standardisation should be included with more care into the project proposals. Later it is difficult to a) have all partners agree upon common rules b) implement metadata capture a posteriori into software. If the development of internal project standards is inevitable, it should take place and be delivered as early as possible.
- With respect to metadata references to external documents like definitions or international standards, the use of Handle PIDs is preferable over urls and other less stable pointers.
- More effort should be made to agree on open legal standards at the beginning of a project.

A more detailed view on the results can be found below in Chapter *Conclusions*.

# 2 The Presentations

For reference, all presentations can be found and downloaded from the workshop webpage at https://verc.enes.org/ISENES2/events/final-is-enes2-workshop-on-workflow-solutions/jwsmg. In this section the speakers and their presentations are given. The content is summarised as far as it concerns handling of metadata.

## 2.1 Session 1: Metadata generation during experiments




### *2.1.1 F. Toussaint (DKRZ): Introduction to Metadata Generation and view back to the previous workshop*

After the general introduction by Kerstin Fieg and Reinhard Budich, Frank Toussaint summarised the result of the first of this pair of workshops *(1st Workshop on MD Generation During Experiments, 2014-01-21/22, Hamburg)*. For MD capture it had as main results:

- A dedicated MD database (MDB) is considered best practice
- There are mainly two forms of MD capture
  - Capture during and integrated into the workflow
  - and MD collection (e.g. from file headers) after data production
- The decision between these two is primarily social rather than technical: The collection after data production is often cleaner and clearer regarding the workflow, liabilities, and permissions.
- Depending on the local workflow and work structure, more than one MDB might be needed, e.g. for data description (classical), for workflow control, for external MD. This again will facilitate metadata aggregation by a clear separation of concerns.

Different classes of metadata capture exist. They all are important for exercises such as CMIP6, primarily:

- harvesting from data files or
- harvesting from workflow systems and
- manual input via forms/DB or via setup files (see figure).

The aim is to minimise the latter, human intensive task of manual input.

For Quality Control MD needs to be checked for

- Completeness
- Consistency,
- Agreement with a requirement which is well documented and provided in good time
- Existence of annotations to document any convergence from standards

Unlike the quality of metadata, the quality of the data content depends on the intended data usage: e.g., physically impossible values are interesting for a model analyser whereas they are not wanted for users of climate projections. Thus possible checks of data values by downstream institutions should be regarded as warnings to the data producer, not as errors. The warnings are to be communicated to the data user as in many cases only there can be decided on this aspect of quality.
Standards need to be enforced so that the delivery of non-compliant data runs the risk of rejection.
Following on from the first workshop, this second workshop was requested to answer the following questions:

- What have we learned after the first workshop?
- Where are we now?
- What are the next steps to ease MD caption for climate model data?

### *2.1.2 B. Lawrence (NCAS, remote talk): ES-DOC – why, what, and where are we?*

Bryan Lawrence gave a talk on the status of the ES-DOC project which is the successor of METAFOR in the development of the Common Information Model (CIM) to describe simulations, ensembles, numerical properties, and other features around model output data. Indeed, the CMIP5 project produced 6.9 Petabytes using 46.3 million compute core hours (as given by the ENES HPC Task Force and Marie-Alice Foujols). To document the products of these efforts CIM gives concepts, dependencies, and data structures of which many were explained – by both: definitions and graphical overviews.

### *2.1.3 S. Wilson (NCAS-CMS): Decorating code to expose algorithmic descriptions of the code to CIM*

The following talk by Simon Wilson was about the challenge to consistently document models which consist of a number of components, prognostic variables, physical parametrisations and the flow of data through the various schemes. In addition, a standard way to document this is needed as, e.g., the CIM which already offers descriptive structures for variables and software components. To describe this, the structure of the model code is kept but direct commenting of code by model experts is still required. To generate the code documentation, the automatic code documenting system ROBODoc was chosen. Specially formatted documentation headers are extracted from (Fortran) source files and stored in a new file whereas metadata can be configured by the user. This provides a new dimension to automated metadata capture relating directly to the model design.

### 2.1.4 A. Gupta (Uni Bergen): Overview of the archiving system at UIB/UNI

The next talk was on the archiving system of the University in Bergen, held by Alok Kumar Gupta. He showed that many journal articles about scientific data agree that there is need to archive scientific data. However, research is held back by the absence of sustainable archives for data and funding agencies, professional societies, as well as publishers each have unfulfilled roles in archive design and data management policies. Alok Gupta then gave a precise overview on the different Norwegian scientific data archives and present and future functionalities, dedicated to the five different types of users: creators, other contributors, data managers, rights holders, and access users.

### 2.1.5 M. Greenslade (IPSL, remote talk): ES-DOC: CIM 2 & CMIP6 – From definitions to specializations

Mark Greenslade provided a more detailed description of the ES-DOC activities for the CIM-2 use in CMIP6. He pointed out, that the CIM data model is partitioned into packages, each package addressing a particular documentation/problem-space. In addition, there is an eco-system of tools & services built upon the data model. At the heart of the ES-DOC eco-system is the python client pyesdoc – created to transform, e.g., a spreadsheet into archived CIM 2.0 documents and those into html pages. This ecosystem has been developed as a result of the lessons learned in CMIP5 and issues discussed within this and other IS-ENES work packages and will be used by sites to facilitate the integration of meta-data publication to local workflow and meta-data solutions.
The presentation followed a guided tour through the wide fields of the CMIP6/CIM problem space. As the CIM structure is static, as far as possible standards are needed. However, finally, the es-doc process will go on to be an evolution, not a revolution, which is the most appropriate practice for an international effort, anyway.

## 2.2 Session 2: Workflows and metadata generation in the context of CMIP6

### 2.2.1 D. Hassell (NCAS), M. Greenslade (IPSL): Automated documentation of CMIP6 simulations from ESGF datasets

For the ES-DOC initiative Dave Hassell discussed various types of simulation documents and their relation to the documentation of a simulation: auto generated descriptions of ensembles, members and simulations, as well as hand generated documentation of performance and of the machine. In addition, descriptions of ensemble axes can be auto initialised but will need human input.
Within this frame, the cdf2cim and pyesdoc libraries have been created and are available in the python package index (pypi.python.org). However, before tests can start, the integration into ESGF stack has to be done as well as some practicalities of converting raw descriptions to CIM2 documents.

The next steps then will be a more detailed rendering of the further info URL which is a key property of the simulation documentation. It describes the location of a web page to use as a starting point for finding out more facts about the simulation. Furthermore some automatic linking in the CIM2 document space has to be supplied.

### *2.2.2 S. Sénési (CNRM / MétéoFrance); replaced by Marie-Pierre Moine (CERFACS): Directly driving data and metadata generation by CMIP6 Data Request content thanks to XIOS*

In place of S. Sénési who was unable to join us, Marie-Pierre Moine gave his talk on dr2xml, a tool to automate the configuration of XIOS-enabled models like NEMO. Some of the main advantages of this tool are that it prevents from stop/restart operations by dynamically analysing the simulated period (and adapting output configuration consequently) and that it avoids the manual configuration of climate models as well as the CMORisation step in the data production workflow as the XIOS written files are CMIP compatible.
Envisaged are further features (automatic configuration of) like spatial regridding (horizontal and vertical), various temporal and spatial aggregations ('cell_methods') and a homemade list of output variables.

### *2.2.3 P.-A.Bretonnière (BSC): Online metadata generation through CMORisation*

After a tour through Barcelona's computing centre, P.-A. Bretonnière gave an outline on online metadata generation based on CMOR3, the latest version of the Climate Model Output Rewriter. Different metadata sources from workstation and HPC environment are merged and written into the CMORised output files which comprise the complete metadata. This allows for easy metadata capture later.
Future plans are to add ES-DOC into the autosubmit part of the workflow, to modularize the CMORisation process, and to add to the metadata a complete history of file processing all along its life to keep track of the changes.

### *2.2.4 J. Durachta (GFDL): Performance Analysis of Chaco: The Next Generation GFDL Workflow Infrastructure*

The last talk of the first day was given by Jeff Durachta on Elements of Workflow Performance Analysis. After an overview of the workflow at GFDL he introduced the Flexible Modelling System (FMS) and the FMS runtime Environment (FRE). This was followed by extensive performance analyses shown in various graphs.
This talk already referred to the theme of the second phase of the workshop which aimed at workflows and not anymore on MD.

## 2.3 Session 2a: Site reports on CMIP6 workflows

The sessions of this second day were not dedicated to metadata generation from experiments but to workflows in general. The presentations are only summarised as far as they refer to metadata handling.

### *2.3.1 J. Walton (MetO): Climate data dissemination using workflow systems*

The report from the Met Office from Jeremy Walton and Mark Elkington started with the MD repository development, where the CREM DB had been had been redeveloped as CREM-2 under the influence of lessons learned from CMIP5. It is based on MySQL, reflects the CIM2 schema, offers a python API, and also covers the management of data processing.
Some coupling to the CIM2 model was implemented. Automation for more complex operations was added.
The dreq (Data request) tool to capture MD from the data producer was described which together with a python wrapper for CMIP6 takes the place of CMIP5's spreadsheets. For the project configuration the general MD schema was evolved to project-specific schemas. Instances can be downloaded in JSON format. Further information was given on the capturing of details from simulation and model.
The publishing of MD relies on an interface to ES-DOC/CIM. Here the CIM2 document schema is used including a project specialisation which is generated from the CREM MD via python and ES-DOC's pyesdoc API.
Finally, some current issues were reported which remain for the future. To them belong the topics required content and coordination within CMIP. For CIM2 publishing the question of referenced vs embedded (xml) documents was discussed.

### *2.3.2 C. Kadow (FU Berlin): A Hybrid Software Infrastructure for Standardized Data and Tool Solutions on HPC within the CMIP6 context*

Freva is a hybrid software infrastructure for standardized data and tool solutions within the CMIP6 context. It is used for evaluation of climate model forecasts, hindcasts, and projections. Here the metadata of analyses' results is stored in a MySQL DB. Some further information on Freva can be found on the MiKliP website at https://www-miklip.dkrz.de .

### *2.3.3 M. Stockhause (DKRZ, remote talk): Improvements in the long-term archiving workflow for CMIP6*

Martina Stockhause described the planned and improved LTA workflow in CMIP6 and deduced various topics to be learned from it. Here the implementation of a furtherInfoURL into the file headers is seen as a big enhancement for CMOR2. It turned out, however, that some main problems are not of technical nature but mainly questions of agreed policies. Other issues communicated in 2014 have been solved as, e.g., by introduction of defined lists of

controlled vocabularies for the DRS, by the collection of data citation information, and by the registration of ancillary MD in the ESGF.

# 3 Conclusions

## 3.1 General aspects – where are we?

Since the initial workshop in 2014, various project partners have implemented MD capture tools into their workflows or they have started to do so. However, because of the size of the challenge, we are far from stable comprehensive systems.

During the course of IS-ENES2, new *workflow metadata*[1] schemes have been developed (see the list of presentations summarised above) which mainly focus on local needs. This is not surprising as models and IO systems are different at the different data producing sites. So the task to convert data and to follow common standards needs to be bespoke to these models. However, it becomes more and more possible to share more of these metadata due to the libraries developed around the metadata and the evolution of more generic IO solutions. The scheme structures of *use metadata*[2] in general use did not evolve to a common scheme but were depending on the specific project. However, many Model Intercomparison Projects (MIP) tend to act in accordance with the global project CMIP. This sometimes requires additional definitions of file header structures in the additional MIPs where these standardisations are not covered by CMIP, e.g., metadata on regionalisation or on climate indices. As a collaboration of IS-ENES2 and CLIPC, however, header conventions for climate indices are developed and are already used by other projects.

Other metadata schemes have been enhanced, e.g., to describe adjacent metadata like experiment, model, ensemble, and run descriptions. Here in the ES-DOC project a strong evolution (see Chapter 2.2.1) led to various new developments regarding the structure of the enhanced CIM data model CIM2. New document types were introduced e.g. to cover the needs of Ensembles (including metadata on axes, axis members, and conformance). In addition, structural considerations on metadata and workflows were done in ES-DOC; libraries like cdf2cim and pyesdoc were the outcome.

So in summary the standardisation is most advanced for the big international projects, e.g., CMIP and for the use metadata in file headers. It is less for minor projects and for metadata describing the workflow and the post processing.

## 3.2 Cooperation – how to get to more common metadata structures?

During the last years, software packages for metadata extraction and control have grown in size and sophistication. This especially holds for metadata capture from file headers.

[1] Workflow metadata describe the workflow of the data production. In an automated workflow they are used for workflow management.
[2] Use metadata are metadata which are essential for the use of the data like unit, coordinate systems, and topic of the data, unlike information on, e.g., provenance, data producer, and scientific effort to which the data belong. Use metadata often are stored in the file header to keep them close to the data.

However, the schemes to store the data in are not yet fully standardised, nor are the file headers. As MIPs become more complex, metadata standards need to be further developed.
On the other hand, there is more and more need for re-use of software for handling and evaluation of scientific data. For this purpose, standardisation and documentation of metadata structures and contents is strongly needed, as well as a homogenisation of the access rights. There are various fields where this should be put more in the focus of scientists, projects, and funders: concerning data and metadata formats, metadata sets including their structure and controlled vocabularies, and legally for, e.g. copyrights.
There is broad agreement that standardisations are important and are of great help for data and software re-use, as well as for metadata capture from file headers. Actually, they are the basis of interchange and re-use.
Why it is so difficult to find common standards? Firstly one should see that this kind of work is not very prestigious. So it is not prioritised. Secondly it needs a lot of communication with partners inside and outside the community to make sure that the outcome is widely accepted. This is corresponding with the fact that standardisation work is not innovative – so it's barely funded. For big global projects like the MIPs a sustained and sensibly funded service to provide the standards and associated tools is needed. Only a project especially dedicated to standards will be able to boost these developments.

### 3.3 Recommendations – where to go?

Looking ahead to find the next steps for metadata handling, one can draw some recommendations from the following IS-ENES2 reports:

- Milestone 4.2: Initial workshop on meta-data generation during experiments,
- Milestone 4.6: Final workshop on meta-data generation during experiments,
- Deliverable 5.3: Basic data access protocols and data quality control.

The main results of the first workshop (MS 4.2) are listed above in Chapter 2.1.1. They have been confirmed and will not be repeated here.
In addition, the following recommendations can be put forward:

A) For use metadata

- Put use metadata into the file header. In general the file header is a good place to put them, as they are at hand whenever the data is.

B) For further information like metadata on model, platform, experiment

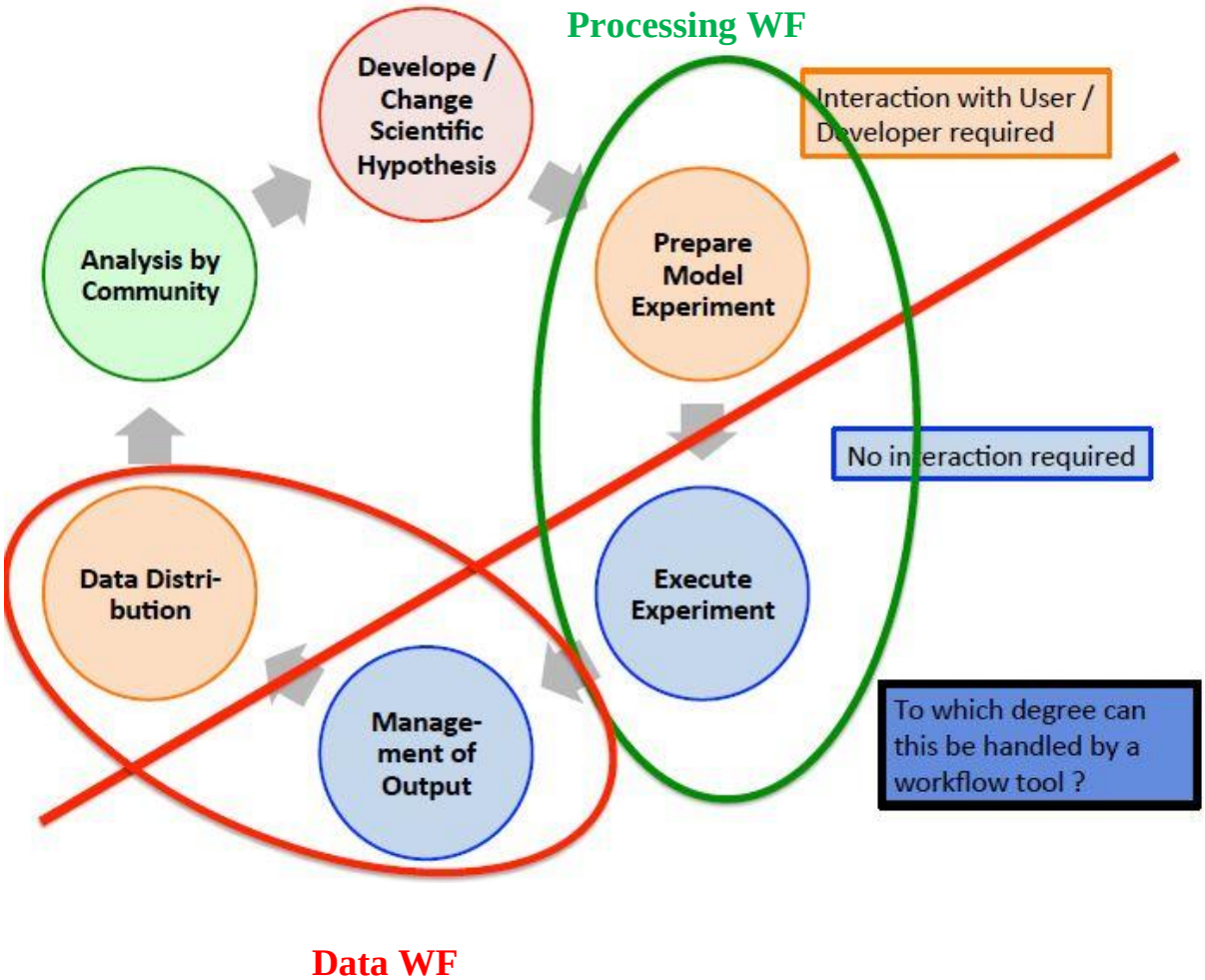
- Try to avoid using URLs pointing to further information. Use Handle PIDs where possible.

C) Standards

- Try to structure your metadata in any project according to existing standards wherever possible. Include this work in project proposals – not just as deliverable papers but as interfaces to be programmed.

- Look in your community for existing content standards. Use them wherever it makes sense.
- Where global standards are not available, use community standards. Where community standards are not available, use project standards.
- Include interfaces to existing standards, e.g. ISO, in project proposals wherever it makes sense – not just as deliverable papers but as interfaces to be programmed.
- Let interface definitions start on the lowest level: For characters: ASCII, utf, ISO; for numbers IEEE754, Hexfloat etc.

*Figure: As the presentations have shown, at various sites parts of the workflow are already automated. The coming years will show, how many other steps can be integrated into the workflow and can be handled by tools.*
*The green and red ovals refer to the Processing Workflow and the Data Workflow parts. The red line rules out a part of the whole workflow which can be practically largely automated.*



- Legal standards of use and copyright for the produced data and software should be part of any project proposal to be agreed on before project start.
- Use open legal standards to keep bureaucracy at a minimum.
- See sustainable funding for standards and associated tooling that supports a world-wide community.

For the collection of metadata on the workflow it may be advisable to write rich log files during the run and generate database input from them after the output data has been approved.

Summarised, the outcome of the workshops was that for work with and homogenisation of metadata, this work should be included with more care into project proposals. Later it is difficult to have all partners agreed to common rules.
This inclusion can be a priori by fixing the standards to use in the proposal. These standards mainly will be external – existing community standards should be used wherever possible and the broad variety of legal standards allows an early fixing as well.
It also can be a posteriori by putting this into work packages to be solved at the very beginning of the project. This is also true for extensions and adaptations to existing standards. To use existing standards as in the former case wherever possible is preferable over creating new ones, as in the latter. Indeed, a project's internal standard does not deserve this name – it's merely a rule.

Furthermore, robust standardisation is needed for metadata content, structure, and formats. With respect to external documents, the use of Handle PIDs is preferable over urls and other less stable pointers.

### 3.4 How far can we go – next steps?

At various sites parts of the workflow are already automated. This includes automated collection and storage of experiment setup, workflow metadata, results of quality checks, and information on post-processing. When we look at the workflow picture in the previous chapter which was taken from the introductory talk of Kerstin Fieg this mainly refers to experiment execution and output management. However, it probably is preferable to setup the Experiments via a configurable workflow tool which is already in the *Prepare Model Experiment* part of the workflow. This can ensure that these data are safely stored in a database or at least in the file headers. This would make more sense than to extract them from the workflow logs later.

However, it remains open, how far Data Distribution and Preparation of the Model Experiment can be automated. As suggested in the picture, these parts of the workflow will probably be the next targets to undergo automation.

# 4 Glossary of abbreviations

| | |
|---|---|
| AGU | American Geophysical Union |
| BSC | Barcelona Supercomputing Center |
| CDNOT | CMIP Data Node Operation Team |
| CEDA | Centre for Environmental Data Analysis |
| CERFACS | Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique |
| CF | The NetCDF standard for climate and forecast data |
| CHARMe | A project for sharing knowledge about climate data by annotations |
| CIM | A Common Information Model for model description MD |
| CLIPC | A Climate Information Portal for Copernicus |
| CMIP5, CMIP6 | Coupled Model Intercomparison Project, Phases 5 and 6 |
| CMOR | Climate Model Output Rewriter |
| CNRM | Centre National de la Recherche Météorologiques |
| CNRS | Centre National de la Recherche Scientifique |
| CORDEX | Coordinated Regional Climate Downscaling Experiment |
| CREM | Climate Research Experiment Management DB (MetO) |
| DB | Database |
| DIF | NASA's Directory Interchange Format for MD |
| DKRZ | German Climate Computing Centre, Hamburg |
| DLR | Deutsches Zentrum für Luft- und Raumfahrt |
| DOI | Digital Object Identifier |
| dreq | Data request, tool to capture MD from the data producer |
| DRS | Data Reference Syntax |
| DSA | Data Seal of Approval |
| ENES | European Network for Earth System modelling |
| ES-DOC | The Earth System Documentation project |
| ESMValTool | Earth System Model evaluator Tool |
| EU | European Union |
| EUDAT | The European Data project |
| FU Berlin | Freie Universität Berlin |
| GFDL | Geophysical Fluid Dynamics Laboratory |
| HPC | High Performance Computing |
| IPSL | Institut Pierre Simon Laplace |
| IS-ENES | The infrastructure project of ENES |
| ISO | International Organization for Standardization |
| MD | Metadata |
| MDB | Meta database |
| MetO | Met Office (official Name), formerly Meteorological Office |

| | |
|---|---|
| MIP | Model Intercomparison Project |
| MPI-M | Max Planck Institute for Meteorology |
| NASA | The US National Aeronautics and Space Administration |
| NCAS | National Centre for Atmospheric Science |
| NEMO | Nucleus for European Modelling of the Ocean, a modelling platform |
| NetCDF | Network Common Data Format |
| OAI-PMH | The Open Archives Initiative's protocol standard for MD harvesting |
| QA | Quality Assurance |
| QC | Quality Control, Quality Check |
| RI | Research Infrastructure |
| STFC | Science and Technology Facilities Council |
| WDS, WDC | World Data System, World Data Centre |
| WF | Workflow |
| WMO | World Meteorological Organisation |
| WP | Work Package |
| XIOS | XML-IO-Server |

# 5 Appendix

The list of participants and the programme below refer to the whole combined workshop in Lisbon although this deliverable refers to the metadata part which was in focus during days one and two.

## 5.1 List of participants

The combined workshop had 52 participants. Included are three persons who gave their talk remotely.

| Name | From | Name | From |
|---|---|---|---|
| Florent Lebeau | Allinea | Marie-Alice Foujols | IPSL |
| Domingo Manubens | BSC | Mark Greenslade | IPSL |
| Joan Lopez de la Franca Beltran | BSC | Sebastien Denvil | IPSL |
| Kim Serradell | BSC | Yann Meurdesoif | IPSL |
| Pierre-Antoine Bretonnière | BSC | Joao Pina | LIP |
| Marie-Pierre Moine | CERFACS | Ben Fitzpatrick | MetOffice |
| Dela Spickermann | DKRZ | Dave Matthews | MetOffice |
| Fabian Wachsmann | DKRZ | Jeremy Walton | MetOffice |
| Frank Toussaint | DKRZ | Mick Carter | MetOffice |
| Hanna Sowjanya Motupalli | DKRZ | Oliver Sanders | MetOffice |
| Joachim Biercamp | DKRZ | Asela Rajapakse | MPI-M |
| Kerstin Fieg | DKRZ | Gunnar Gorges | MPI-M |
| Ksenia Gorges | DKRZ | Luis Kornblueh | MPI-M |
| Martin Schupfner | DKRZ | Matthias Bittner | MPI-M |
| Martina Stockhause | DKRZ | Reinhard Budich | MPI-M |
| Pavan Siligam | DKRZ | Sergey Kosukhin | MPI-M |
| Torsten Rathmann | DKRZ | Annette | NCAS |

| | | | |
|---|---|---|---|
| | | Osprey | |
| Björn Brötz | DLR | Bryan Lawrence | NCAS |
| Christopher Kadow | FU Berlin | David Hassell | NCAS |
| Ingo Kirchner | FU Berlin | Rosalyn Hatcher | NCAS |
| Jeffrey Durachta | GFDL / NOAA | Simon Wilson | NCAS |
| V. Balaji | GFDL / NOAA | Grenville Lister | NCAS / UREAD |
| Chandin Wilson | GFDL /NOAA | Hilary Oliver | NIWA |
| Arnaud Caubel | IPSL | Uwe Fladrich | SMHI |
| Claire Levy | IPSL | Ag Stephens | STFC |
| Josefine Ghattas | IPSL | Alok Gupta | Uni Bergen |

## 5.2 Workshop programme

Tuesday, 27 September 2016

| | | |
|---|---|---|
| 09:45 | Introduction and results from 1st workshop in 2014 | R. Budich, K. Fieg |
| **Session 1: Metadata generation during experiments** | | Chair: C. Kadow |
| 10:15 | Introduction to Metadata Generation | F. Toussaint, DKRZ |
| 10:30 | Introduction to the common information model (CIM) | B. Lawrence, NCAS (remote talk) |
| 10:45 | Decorating code to expose algorithmic descriptions of the code to CIM | S. Wilson, NCAS |
| 11:00 | Overview of the archiving system at UIB/UNI | A. Gupta, Uni Bergen |
| 12:00 | ES-DOC: CIM 2 & CMIP6 – From definitions to specializations | M. Greenslade, IPSL (remote talk) |
| **Session 2: Workflows and metadata generation in the context of CMIP6** | | Chair: F. Toussaint |
| 13:30 | Automatic documentation of CMIP6 simulations from ESGF datasets | D. Hassell, NCAS |
| 14:00 | Directly driving data and metadata generation by CMIP6 Data Request content thanks to XIOS | S. Sénési, CNRM / MétéoFrance; replaced by Marie-Pierre Moine, CERFACS |
| 15:00 | Online metadata generation through CMORisation | P.-A.Bretonnière, BSC |
| 15:30 | Performance Analysis of Chaco: The Next Generation GFDL Workflow Infrastructure | J. Durachta, GFDL |

Wednesday, 28 September 2016

| | | |
|---|---|---|
| **Session 2a: Site reports on CMIP6 workflows** | | Chair: K. Fieg |
| 09:30 | The Hamburg CMIP6 Workflow | L. Kornblueh, MPI-M |
| 10:00 | Climate data dissemination using workflow systems | J. Walton, MetO |
| 10:30 | Climate workflow at IPSL: from CMIP5 to CMIP6 | S. Denvil, IPSL |
| 11:30 | A Hybrid Software Infrastructure for Standardized Data and Tool Solutions on HPC within the CMIP6 context | C. Kadow, FU Berlin |
| 12:00 | The CMIP6 ingest-to-publication pipeline at CEDA | A. Stephens, STFC |
| 12:30 | Improvements in the long-term archiving workflow for CMIP6 | M. Stockhause, DKRZ (remote talk) |

| | | |
|---|---|---|
| **Session 2b: Special workflows** | | Chair: R. Budich |
| 14:30 | Workflow for routine evaluation of CMIP6 models with the ESMValTool | B. Brötz, DLR |
| 15:00 | Remote Workflow Enactment using Docker and the Generic Execution Framework in EUDAT | A. Rajapakse, MPI-M |
| 15:30 | General Discussion on WF, MD and CMIP6 | All |
| 16:10 | Keynote: Convergence of computation and data workflows | V. Balaji, GFDL |
| 16:40 | General Discussion on Keynote | All |

Thursday, 29 September 2016

| | | |
|---|---|---|
| **Session 3a: Workflow tools** | | Chair: V. Balaji |
| 09:00 | Keynote: Asynchronicity L. Kornblueh | MPI-M |
| 09:50 | General Discussion on Keynote | All |
| 10:20 | Weather and climate models: preparing development workflows for Exascale | F. Lebeau, Allinea |
| 10:40 | Software stack deployment for Earth System Modelling | S. Kosukhin, MPI-M |
| **Session 3b: Scheduling** | | Chair: M. Carter |
| 11:30 | Keynote: Cylc - Recent developments & future plans | H. Oliver, NIWA |
| 12:20 | Scale and breadth of cylc usage at the Met Office | D. Matthews, MetO |
| 12:40 | Cylc from NCAS point of view | G. Lister, NCAS |
| 14:00 | ESIWACE Cylc development and support plan | D. Matthews, MetO |
| 14:25 | A progress report on the rewrite of the GFDL FMS workflow to use Cylc and discrete tools vs. a monolithic job flow | C. Wilson, NOAA |
| 14:50 | Comparison of autosubmit / cylc / ecflow | D. Manubens, BSC |
| 15:15 | FreVast - combining modelling with diagnostics at university level | I. Kirchner, FU Berlin |
| 15:40 | General Discussion and wrap-up | R. Budich |

# 6 Acknowledgements

*Thanks go to all presenters and organisers of the workshop, especially to Kerstin Fieg, Dela Spickermann, and Reinhard Budich, who did the main part of the administrational and organisational work. Additionally, of course, they go to the reviewers for many useful comments.*