IS-ENES2 and ESDOC:

Actions and Notes from the Abingdon Meeting, February 2015

This document summarises discussions held as part of an IS-ENES2 meeting on cross-cutting IS-ENES2 ESDOC activities and plans, and provides a list of actions together with a timeline to meeting a range of internal and external goals.

ESDOC is an a distributed activity, in terms of the funders, the teams working on the project, the location of services, and with respect to the repositories hosting relevant software. There are many stakeholders. Accordingly, we have presented these notes using the RM-ODP viewpoint mechanism (RM-ODP is summarised in box 1) coupled with (what should be, but isn’t yet, a clear roadmap which describes how we get from where we are to where want to go).

**RM-ODP: The Reference Model for Open Distributed Processing.**

**The key concept is to break a distributed processing problem up into five viewpoints:**

*The enterprise viewpoint*focuses on the purpose, scope and policies for the system. It describes the business requirements and how to meet them.

*The information viewpoint* focuses on the semantics of the information and the information processing. It describes the information managed by the system and the structure and content type of the supporting data.

*The computational viewpoint* enables distribution through functional decomposition on the system into objects that interact at interfaces. It describes the functionality provided by the system and its functional decomposition.

The **e***ngineering viewpoint* focuses on the mechanisms and functions required to support distributed interactions between objects in the system. It describes the distribution of processing performed by the system to manage the information and provide the functionality.

*The technology viewpoint* focuses on the choice of technology of the system. It describes the technologies chosen to provide the processing, functionality and presentation of information.

One advantage of doing this is that it made clear that our discussions did not cover the full spectrum of issues we need to consider to meet our infrastructure objectives.

# The Enterprise Viewpoint

In this section we summarise what we are trying to achieve, by providing the big picture and specific use cases.

**What we’re trying to achieve:**

1. Define standard vocabularies needed for CIM2 and CMIP6 and any other projects prioritised by IS-ENES2.
   1. This is in the context of meeting WGCM requirements.
2. The resultant vocabularies need to be easy to use, and update (both to fix errors and extend).
3. Data driven – from variable to information
4. Meet WGCM requirements in terms of tables, citation, narrative and granularity

*Use cases/requirements for*

* Comparator/viewer etc…
* Link back from data – variable bundles centric view
* Document MIPs via experiment SV
* Included in statement of purpose (BAMS/EOS style) Veronika’s paper

*Flow chart for the modelling groups, explaining what we are expecting them to do*

# The Information Viewpoint

*Standard vocabularies*

* Process and timeline for defining SV
* SV contents
  + Model:
    - questions and answers
    - narrative aspects / abstract
    - Tuning
      * Text below abstract for top level and components
    - Grids
    - Glossary/documentation
    - Where and how to add structure?
  + Experiment info and framework
  + Simulation
  + Forcings (all in the one place)
  + Process for involving the wider community, including WGCM, WGSIP, CLIVAR panels, end users
* Process for managing SV
* Converted SV mindmaps into tables (to present to community)

*What we had in CMIP5*

* Experiment docs (delivered by Charlotte)
* Model docs (delivered via the questionnaire)
* Simulation docs (delivered via the questionnaire)
* Platform docs (delivered via the questionnaire)
* Forcings (delivered in both files and the questionnaire, badly in both cases, but for different reasons, the former because modelling groups didn’t fix errors, and in the latter because our structure sucked).

*Experiment definitions*

* Need to understand forcings
* Be sure experiment structure works for our experiments

# Computational and Engineering Viewpoint

**What tools need to exist?**

*Standard vocabularies*

* SV management tools
  + SV enforcement/validation (core/Tier1,…)
  + SV version control
  + SV server
    - Required: DRS code list common to ESGF/ES-DOC/CMOR.
    - WIP issue (will need to wait to see where things evolve)
  + Can use Google Doc spreadsheet instead of mind map (warning: Google Docs can be very laggy)

*Communication tools*

* Launch of tools
  + Regular demos / animated howto
* Request for help for CMIP6 process
  + Defining SV
  + Tools
* New people join es-doc list
* Use CoG, gitHub
* General communications coord. with WIP

*Notebook*

* Make pyesdoc compliant
* Able to export pdf
* Confirm it can describe the DECK (HiresMIP and Decadal)

*DRS based list of codelists managed in a tool*

* In NERC Vocabulary Server?

*Metadata entry tool (questionnaire)*

* Granularity of questions an issue
* Incentives for modelling groups to complete the metadata entry?
  + Tabular content of metadata
  + Clear examples of what a complete metadata collection looks like

*Metadata inspection/comparison tools*

*pyesdoc*

The pyesdoc library is designed to simplify and streamline documentation creation, validation and publishing. It will be made available for modelling groups to document CMIP6 models and simulation. The capabilities of the library were demonstrated in detail:

* Archival: Documents can be pulled from remote sources and stored in a local file system archive.
* Extensions: Documents are parsed and injected with extra information (extensions) derived from the original document. An example of this is flattened model component hierarchies.
* I/O: Documents can be simply written and read from the local file system. This is leveraged by the archival feature.
* Ontologies: The relationship to the ES-DOC meta-programming framework was demonstrated. This supports defining ontologies, such as Metaphor CIM v1.x, as a set of python functions. The definitions then act as the inputs to code generators that forward engineer pyesdoc code.
* Publishing: ES-DOC hosts a remote web service API. Using pyesdoc an institute can publish to the remote web-service.
* Serialisation: The pyesdoc serialiser supports several encodings such as JSON, XML, HTML …etc. The importance of this is that documents are decoupled from from document encodings.
* Validation: The pyesdoc validator is flexible and extensible. It ensures that documents are valid according to the rules defined in the meta-programming library.

Actions for pyesdoc

1. Controlled vocabulary validation was seen to be a weak spot. This has to be implemented for CMIP6.
2. Need to support PDF documents.
3. The publishing API needs securing at the transport and request levels.
4. Viewer must be extended to handle offline documents.
5. Comparator to handle single component models and simulation forcings.
6. An integrated document search and browse is required.

# Roadmap

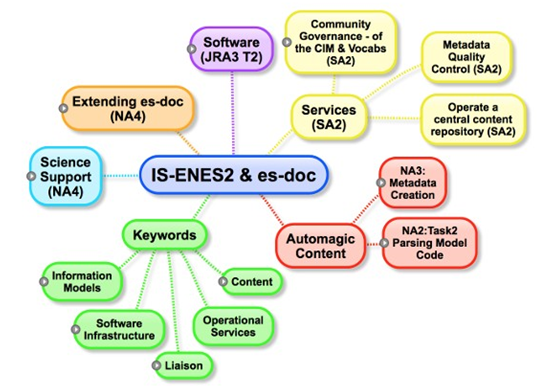
**Timeline (modified from Eric’s slides at Abingdon meeting)**

* **March 2015**: Announce and/or demonstrate CMIP5 tools
* **June 2015**: Final version of CMIP5 metadata (including tables…)
* **Summer 2015**: Experiment/forcings SV draft documents (**Charlotte** to coordinate)
  + Engage with MIPs **March 2015**
  + Page proofing draft documents (prioritising base MIPs) **June-December 2015**
  + Experiment specifications complete **December 2015**
* **Summer/early fall 2015**: Simulation SV process (**David**)
* **Fall 2015**: Grid SV and associated CIM changes (**Bryan**)
* **January 2016**: Models SV 95% complete
  + **April 2015**: Organize original mindmaps into simpler tables for community review - (**David**)
  + **May/June 2015**: WGCM statement of purpose (**WGCM**)
  + **June 2015** : Move from mindmap to process flow
    - **March 2015**: Proof of concept using CMIP5 sea-ice (**Bryan**)
    - **June 2015**: Apply to 7 other realms -
      * Who: ocean (**Eric**), atmosphere (**Charlotte**), atmosphere chemistry / aerosols (**Charlotte**), land surface (**Ksenya**), OBG (**Eric**), land ice (**David** to find someone in Met)
  + **July 2015**: Begin community review of model SV (**ES-DOC PIs/Eric**)
  + **September 2015:** Community review ends (**ES-DOC PIs/Eric**)
  + **November 2015:** Revision to V.1 and documentation
* **June 2016**: Historical forcings
* **Fall 2016**: Models 100% complete

**Tasks from Abingdon meeting (from Eric’s slides)**

* Involve wider community
  + WGCM (discussion with Sandrine Feb 2015)
    - propose paragraph to be sent by WGCM to find/commission 2 names
    - Scope:
      * Review of SV contents & list use cases (what it is for?)
        + Statement of purpose (documentation, new science, …)
      * Organise wider community review when needed
      * Own the groups who own theSV
      * **Sébastien** and **Eric**: send paragraph to Sandrine ( Monday 22 Feb)
  + Timeline: Statement of purpose for June, review of community September 2015
  + Other groups including: WGSIP, CLIVAR panels, “end users”
    - Send them following and ask for feedback (June 2015)
      * Statement of purpose
      * PDF of CMIP5 mindmaps
      * Examples (tables, …)
* Use cases
  + Comparator/viewer etc…
  + Link back from data – variable bundles centric view
  + Document MIPs via experiment SV
  + Use cases included in statement of purpose
* Forcings SV: **Mark Elkington** (with **Tim Johns**)
  + In CMIP5 from files
  + Store in metadata rather than in file
  + How models conform to experiment protocol
* Experiments SV: **Charlotte Pascoe**
  + Use notebook to describe experiments
    - Start by describing DECK + a few other (decadal, HighRes mip) internally
    - Then engage with MIP coordinator
    - Produce PDF to be used by MIP coordinator
  + Demonstrate engage with community + early benefit for CMIP6
* Model SV:
  + Simple questions and answers leading to 2D table - Control what is needed for IPCC tables
  + Have more common standards question for each component – physical processes, algorithms – list them as suggestions rather than fixed form (dropped down list)
  + Documentation and examples
  + **Bryan** **Lawrence** to organise existing sea-ice mindmap in less contrained way: use process flow
  + Organize original mindmaps into simpler tables for community review (e.g. remove conditional,…)
  + Revisit CIM organisation to reflect new structure in viewing tools
* Grids
  + Use cases – produce basic info, including degrading/coarsening
  + Model grid high level information should not be automated (c.f. mindmaps)
  + Data grid info
    - Use files metadata harvest for data files resolution
    - Who: **David** from cfpython
* Simulations
  + Requirement: unique URI or name based on DRS rules?
  + Simulations link model, initial state, ensemble & forcings
  + Determine what can be captured from netCDF headers?
  + Need commentary/note free text section
  + Workflow needed - should be mostly automated – still needs editing capacity
  + Conformance (see forcings) – only flag when not conforming (**Mark E**)
  + CIM reflects ensemble pattern (to be described once)
  + Proof of concept using CMIP5 (**David**)

**Activity Packages (related to IS-ENES2 WP – text edited from** <https://redmine.dkrz.de/projects/is-enes/wiki/Isenes2esdoc_/6>**).**



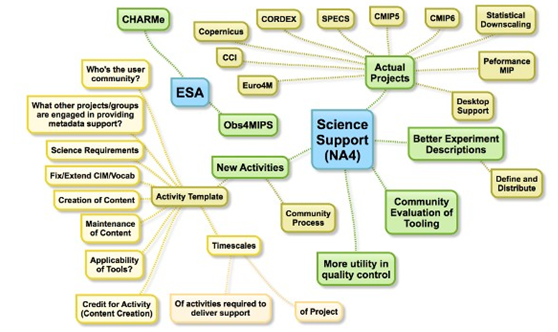
**(WP3) NA2: Task2 Model Structure and Code Evaluation (UREAD‐NCAS)**

The key aim of this sub­task is to prepare the ground for subsequent community discussion of the different approaches in use in Europe. The first activity will be to establish a consistent methodology for documenting key model components, the exchanges between them, and their scaling properties ­ across all the major European models. This methodology will exploit the model metadata work in other work packages, and the service documentation activities. It is likely that the resources available will only allow the initial analysis of two models, but this will then feed into a workshop to establish the efficacy for more wide usage of the methodology by model groups. This workshop will also be used to help focus work on computational cores (subtask 2.2) and programmability (subtask 2.3).

**(WP4) NA3: Task 3. Metadata creation and usage. DKRZ, MetO, CNRS‐IPSL, UiB**

This networking activity will promote the sharing of experiences and designs in the area of meta­data capture in the ESM experiment process through two workshops organised by DKRZ. The aim will be to encourage investment in software and working processes that will allow more comprehensive meta­data to be collected more efficiently. Further, the development of workflow and diagnostic solutions will be influenced by the meta­data requirements. To support the workshop, the Met Office and DKRZ will develop documents that identify key interfaces between the meta­data and the experiment definition and modelling processes, and explore design solutions.

**(WP5) NA4: Task 2: Meta‐data, interoperability and standardisation STFC, CERFACS, CSAG, MF‐CNRM, UNIMAN, CSAG) in collaboration with ESA**



This task focuses on extending the FP7 METAFOR “Common Information Model” (CIM) to support new data and activities, better experiment description, more utility in quality control – and coordinate generation of new content. CIM extensions and new controlled vocabulary will need to be coordinated on a global scale, and notably in direct coordination with the US­led CURATOR and National Unified Operational Prediction Capability (NUOPC) consortiums.

This task will establish and manage a community process for identifying and delivering CIM upgrades to improve and extend support for global climate model documentation, including better descriptions of coupling and frameworks, and statistical downscaling methodologies. New support for regional climate models and re­analyses data will also be initiated. This will lead to new CIM controlled vocabularies (CV[[1]](#footnote-1)) to be handed to SA2 for community governance. Working groups will be established for each theme, and these will, in the first 6 months, identify and establish contact with the key community projects (in Europe and elsewhere) in each area. Liaison with similar projects working on remote sensing and ground based observations will be important – particularly in the context of GMES and WMO Climate Service Information System. Close collaboration will take place with ESA to foster the exploitation of European satellite data archive and the new generation of missions (Earth Explorers & Sentinels) and products (in particular from ESA Climate Change Initiative (CCI) when available). More precisely, given the direct involvement of ESA within IS­ENES2, we will ensure interoperability of the ESA CCI database repositories. The scientific coordination and development of CIM and CV extension will be done in NA4 whereas the technical support for the resulting governance activities will be covered in SA2.

NA4 will further organise the community evaluation of CIM generation software, developed both in JRA3 as well as in other projects. A gap and duplication analysis as well as a sharing of best practice and experience with various experimental generation of CIM content will be performed.

*Extending use of Metadata*

Metadata providing detailed and structured information about data holdings is fundamental to the archive infrastructure. Common ontologies and vocabularies allow different parts of the federation to interact in ways that are transparent to the user. The "Common Information Model" (CIM) developed by the FP7 project METAFOR to describe global climate models and simulations will be extended to support more data types, such as regional models (CORDEX), statistical downscaling, re­analyses and observations (see also task2 re re­analyses)

Networking activities will be essential for collaborations with key US groups as well as other related initiatives such as Eranet CIRCLE­2, FP7 Euro4M, G8­ExArch, and relevant European research and Integration projects.

These changes will also make it possible to use the CIM to define and distribute experimental plans in a consistent manner, and use the CIM as a critical input to automating data management functions such as quality control.

The standardisation efforts will be extended to the CORDEX programme statistical downscaling component, with specifications of the experiment design and data format.

Regarding the development of metadata standards, the scientific coordination and development of CIM and CV extension will be done in NA4, coordinated with SA2, which will provide the technical support for the resulting governance activity. This task will establish and manage a community process for identifying and delivering CIM upgrades. This will lead to new CIM controlled vocabularies (CV) to be handed to SA2 for community governance.

**(WP7) SA1: Task 2: Services on European ESMs**

**MetO, KNMI, SMHI, CNRS‐IPSL, Met.no, UiB, CMCC, MPG, MF‐CNRM**

Different levels of services will be offered depending on the experience of the different ESM groups to support external users. All services described will be available since the beginning of the project.

At level 1, the documentation established during IS­ENES1, under free format and under the CIM metadata format, will be maintained up­to­date and accessible to all through the ENES portal. A contact person will also be identified to answer specific questions about the ESM when needed.

This level of services will be offered for all European ESMs used to run the CMIP5 simulations, helping the CMIP5 data users to better understand the characteristics of the ESM that produced the data.

**(WP8) SA2: Task1: Core Data Services**

**Activity: Quality Control Service (DKRZ , STFC)**

This service activity supports and assists the data and metadata quality control. Starting point will be the quality control process which has been coordinated and examined for CMIP5. Improvements are expected from corresponding NA and JRA activities.

**(WP8) SA2: Task 2: Meta‐data Services**

New developments in the metadata part for CIM related tools are integrated into the metadata services and will become part of the services activities during the course of the project.

Meta­data services in SA2 are focussed on support of meta­data access while Meta­data services in SA1 are concentrating on population of CIM instances.

*Activity CIM Governance (STFC)*

This task organises the maintenance of CIM schema and related controlled vocabularies with emphasis on model and experiment descriptions. Outputs from NA4 metadata networking task in the form of scientific requirements from CORDEX and the impacts community will be integrated.

*Activity: CIM Repository (DKRZ, CNRS‐IPSL)*

This task operates the model and experiment metadata repository based on CIM instances from the METAFOR project and CMIP5. Improvements in CIM metadata services are closely related to developments in NA4 and JRA3.

**(WP11) JRA3 Task 2: Meta‐data Services Package**

**CNRS‐IPSL, UREAD‐NCAS, DKRZ, MetO**

The main aim of this task is to build the meta­data creation and manipulation tools needed to support the archive infrastructure. The schema defining the CIM will be updated in NA4, this task will create the tools needed to support operational use of the CIM.

*2.1 Tools for metadata capture and generation (UREAD‐NCAS, MetO)*

The primary tool used thus far for meta­data creation (in the context of IS­ENES support for CMIP5) has been the METAFOR questionnaire. This tool consists of a graphical user interface handling answering hundreds of questions about model capability and configuration. The existing tools are heavily customised for CMIP5, and need re­factoring to improve flexibility for alternative users. As CIM capability is extended within NA4, the new tool will be modified accordingly, and deployed in support of SA2. The MetO internal tool for meta­data will also be modified to directly generate CIM content so that alternative approaches to meta­data entry can be compared.

It's not obvious that we need to continue to invest in the CMIP5 questionnaire, unless we need it to support other projects, and the esdoc questionnaire is not available, or the cimnb is not appropriate. If we don't, then it would be sensible to move the CIM development in response to requirements from NA4 to this activity, with no change in funds, since the science support role has grown rather larger.

*2.2 Tools for metadata services (including repository) (CNRS‐IPSL, DKRZ)*

The existing meta­data repository is based on XML1 storage, and is not thought to be fit for purpose. A revised storage back­end will be constructed, perhaps based on JSON1 storage artefacts, and the web service interfaces extended accordingly (XML for transport will still be supported, but additional JSON31interfaces will be provided to facilitate CIM navigation by web browsers). Improved support for data quality records and annotation will be provided.

**Who does what?**

* See timeline (above) for names attached to tasks.

1. Also known in this document as Standard Vocabularies (SV) [↑](#footnote-ref-1)