CMIP Licensing and Access Control

V1.0 WGCM Infrastructure Panel

Author: B.N. Lawrence

Executive Summary

The existence of, and content in, the CMIP5 terms of use are well kept secrets (insofar as it's hard to find them if you think you are after a data license via google) and if you don't know they exist, you are unlikely to know their content. Nonetheless, it is the requirements of these terms of use, and in particular the more restrictive one that provides a significant amount of the reason why ESGF has a complex access control paradigm – so that the data holders can respect both the open access mandate of some partners, and the restricted access mandate of others. (The other main reason for the ESGF access control is to avoid denial of service attacks via spurious download requests.).

Some caches of the ESGF archive hosted at servers that are not part of ESGF provide data to users without in any way imposing the terms of use requirements, which suggest that the community is at best honouring them in spirit, and likely not even that.

It is proposed here that a change to the way the terms of use are both described and implemented for CMIP6 and successors would both make their provisions and restrictions more clear to all data users, and would make it easier to build, maintain, and use ESGF access control.

The proposal is that (1) a data license be embedded in the data files, making it impossible for users to avoid having a copy of the license, and (2) the onus on defending the provisions of the license be on the original modelling centre, not on ESGF (or any other downstream data provider).

Introduction

This document proposes a new way of imposing CMIP terms of use – embedding a digital license in the netCDF headers of CMIP5 data. The proposal is introduced after first describing the background to why such a proposal would both improve user exposure to the terms of use and make it easier to build ESGF services.

Background

CMIP5 data are made available under a "terms of use agreement"[1], which states, amongst other things, that

> Users registering to access CMIP5 output will be granted access to some or all of the data, depending on which of the following terms of use are agreed to:
>
> Terms of use for data limited to non-commercial research and educational purposes
>
> a) I agree to restrict my use of CMIP5 model output for non-commercial research and educational purposes only.
>
> Results from non-commercial research are expected to be made generally available through open publication and must not be considered proprietary. Materials prepared for educational purposes cannot be sold. These restrictions may only be relaxed by permission of the individual modeling groups responsible for the simulations.
>
> OR
>
> Terms of use for unrestricted data:
>
> b) I understand that the subset of CMIP5 model output that will be made accessible to this group has been designated for "unrestricted" use.

All but four modelling groups make their data available in an un-restricted way.

There are additional requirements (see appendix) on the users of the data (whether restricted or not), and additional requirements on those who chose to provide a local archive and redistribute. These additional requirements were added after some partial replica sites were established, and since they were added only one group has specifically signed up to them.

From a user perspective, the terms of use are essentially about making sure that the users respect the usage requirements (if they exist), cite the data appropriately, understand the limitations of the data (and effectively provide the data provider with evidence the user understood the data had no warranty of fitness of purpose), and will provide feedback on the data and their usage. They also are required to agree with this statement: "*I understand that I may not redistribute the data more widely without abiding by additional terms of use enumerated below*." Those additional terms appear in the appendix to this document, but they were primarily about ensuring that redistributors themselves passed on the terms of use, that the redistribution archives were kept as close as possible to the "official" archive contents, and that statistics of usage could be collected by ESGF (for the WGCM and others).

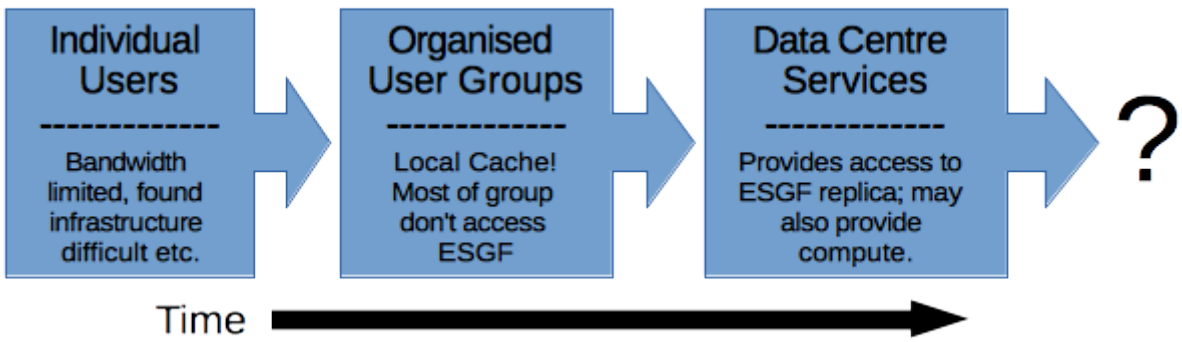
From an ESGF point of view, ensuring that users were subject to the terms of use meant the development, deployment, and management of a complex

[1] CMIP5 terms of use are available here: http://cmip-pcmdi.llnl.gov/cmip5/terms.html (August 06, 2014)

distributed authentication and authorisation system. The system eventually constructed allowed PCMDI to control the authorisation of all users and to propagate those attributes to all sites anywhere in the ESGF federation so those sites could implement access controls which respected the centralised authorisation. The system supported both browser-initiated downloads, wget scripts, and opendap through the netCDF interface (which required modifications to netCDF itself, which Unidata provided in the trunk for netCDF). While it is fair to say that the system provided many users with a lot of grief, it was a major technical achievement to deliver such a system at all, and it was significantly better than the alternatives which were discussed – which all had to meet the criterion of combining centralised authorisation with decentralised authentication.

The criterion of centralised authorisation and decentralised authentication followed from the (perceived or real) requirement to ensure that all users were exposed to the terms of use, and to scale into the distributed environment. An additional significant requirement was to provide a hurdle which enabled the data providers to manage and/or avoid denial-of-service by repeated and/or large downloads inadvertently or by malicious parties.

The registration system was also envisaged to allow the collection of statistics as to data usage, and a route to contacting users if/when data issues were encountered. In practice however, it was not used for notification, and the data usage stats have become less useful, because in practice what has actually happened is depicted in this diagram:



The trend during CMIP5: *from download and process at home, to using services!* Early adopters downloaded locally, got frustrated with bandwidth and tooling, and started automating. Groups rapidly self-organised and created their own cache and relied on a couple of local experts to obtain ESGF data. Bigger groups (including some ESGF members) began to manage proper replicas and put local computing alongside. Given data volumes we are likely to see national archives and facilities for CMIP6. *(Figure concept: Stephan Kindermann, DKRZ.)*

The number of "dark users" (served by local caches) and data centre services is unknown, although where data centre services were provided by ESGF sites like BADC, users still need to register with PCMDI to use the local cache. Many

if not most dark users will never have been exposed to the terms of use. An additional consequence of this trend is that the concept of "data download volumes" being a meaningful metric of usage has diminished (since many, if not most, users are accessing the data directly via a file system).
So the status quo is that we have an access and authorisation scheme that is complex, expensive to maintain and deploy, and is not actually meeting the goal of ensuring users see (and comply) with the terms of use. (Of course it could never have ensured compliance!). Many communities have bypassed ESGF completely. It no longer provides accurate usage statistics. It does however ensure that some of easier denial-of-service attacks are precluded (but, by having a central authorisation point at PCMDI, have some significant weaknesses[2].)
Even the terms of use themselves are not legal documents and can really only be considered as "requests for compliance". While from a WGCM point of view that's probably acceptable, it is a problem both for the modelling centres who do want to constrain some aspects of access, and worse still, it actually leaves the data providers (ESGF) in somewhat of a legal limbo. In the UK at least, the ESGF data providers could be held legally liable for misuse of the data by users, since ESGF does not have a legally valid license waiver in place. While US government entities are immune from prosecution, that would not protect the other US data providers. Institutions in other nations probably also have similar issues.
The question then is: Can we do better? The answer is yes, and the good news is it should be relatively easy to do (certainly easier than what we have done thus far).

Data Licensing

People often confuse the practice of providing "open access" for <x> with "putting <x> in the public domain" in terms of intellectual property. In fact, in most jurisdictions it is necessary to provide a license which releases claims on IP to ensure users are free to use <x> as they wish. This is the spirit in which the existing CMIP5 terms of use make it clear that most of the data can be used in an unrestricted manner.
In licensing digital property it is necessary in most jurisdictions to ensure that prospective users are **most likely** to have seen, or know of, the appropriate license under which they use that digital property. It is this which motivates

[2]When PCMDI did have a major hardware outage, it was possible with heroic efforts by some individuals for other organisations to take over key centralised services during the outage – but it was far from automated, and very time consuming to make happen. There had never been any funding (or will?) for hot-backup for the centralised services.

the existing CMIP5 requirement that ESGF users click through the terms of use – but of course that only applies to original downloaders. Those downloaders are **supposed** to minimise their sharing, but we know that this doesn't happen in practice. What then is the solution?

The easiest solution requires two steps: (1) to embed the terms of use in the data files themselves, and (2) put a banner on all ESGF and derivative sites notifying users that the terms of use are embedded in the files. Of course (2) becomes more difficult to ensure in derivative sites, but if the terms of use themselves state the requirement in the files, it is far less likely that downstream sites can claim to be unaware of the constraints.

The embedding can easily be achieved by simply making the license one of the required file attributes (and ensuring that it appears via CMOR).

This approach would have a two significant benefits: firstly, **all** users of the primary data are exposed to the license and **most likely** to have seen it, and secondly, by removing the requirement to click through, it changes the requirement on ESGF in such a way that it opens up many more technical solutions to user management and denial-of-service.

If we go the step of embedding the terms of use in the files, we can go one step further, and turn the terms of use into two components: something more approximating a legally valid license which provides the proper legal waivers as to use and quality of the data, and a set of "requests for compliance" as to citation etc. The first component could exist in two variants, the restricted and unrestricted ones.

We use the phrase "something more approximating" because there is no unique licensing solution that is known to work in all jurisdictions. However, we believe that the proposal to follow is the best available.

Specific Proposal

1. All CMIP6 (and later) data files held in the ESGF should exploit either one of the following variants of the creative-commons licenses, by including exactly one of these texts in the headers:
    - CC-BY-SA: CMIP6 Model Data produced by <Your CentreName> is licensed under a Creative Commons Attribution "Share Alike" 4.0 International License (http://creativecommons.org/licenses/by-sa/4.0/). The data is hosted via http://esgf.org. Permissions beyond the scope of this license may be available at http://pcmdi.org/cmip6/terms-of-use.
    - CC-BY-NC-SA: CMIP6 Model Data by <Your CentreName> is licensed under a Creative Commons Attribution

"NonCommercial Share Alike" 4.0 International License (http://creativecommons.org/licenses/by-nc-sa/4.0/). The data is hosted via http://esgf.org. Permissions beyond the scope of this license may be available at http://pcmdi.org/cmip6/terms-of-use.

2. The headers of the netcdf file should also include a request for data users to register for updates about their data, which would provide a metric of usage, and provide a means of alerting them in the event of a retraction or update.
3. The pages at http://pcmdi.org/cmip6/terms-of-use should be modified to include the full text of the license, and include the more expanded and suitable waiver from the text of the CC license (section 5 of http://creativecommons.org/licenses/by/4.0/legalcode). This waiver may also be included in the NetCDF headers (to be decided later).

Appendix One: Complete Terms of Use for CMIP5

(As downloaded, August 05, 2015)

Terms of use agreement for CMIP5 model output

All model output in the CMIP5 archive is available for "non-commercial research and educational purposes." A subset (about three-quarters of the models) of the data has also been released for "unrestricted" use, see table in the document "Modeling Groups and their Terms of Use".

Users registering to access CMIP5 output will be granted access to some or all of the data, depending on which of the following terms of use are agreed to:

Terms of use for data limited to non-commercial research and educational purposes

- a) I agree to restrict my use of CMIP5 model output for non-commercial research and educational purposes only.
  Results from non-commercial research are expected to be made generally available through open publication and must not be considered proprietary. Materials prepared for educational purposes cannot be sold. These restrictions may only be relaxed by permission of the individual modeling groups responsible for the simulations.

OR

Terms of use for unrestricted data:

- a) I understand that the subset of CMIP5 model output that will be made accessible to this group has been designated for "unrestricted" use.

For both groups of users, the terms of use include these additional statements:

- b) I will hold no individual(s), organization(s), or group(s) responsible for any errors in the models or in their output data.
- c) In publications that rely on the CMIP5 model output, I will appropriately credit the data providers by an acknowledgment similar to the following:
- "We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table XX of this paper) for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and

Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals."

- where "Table XX" in your paper should list the models and modeling groups that provided the data you used.
- d) I understand that Digital Object Identifiers (DOI's used, for example, in journal citations) will be assigned to various subsets of the CMIP5 multi-model dataset, and when available and as appropriate, I will cite these references in my publications. I will consult the CMIP5 website (http://cmip-pcmdi.llnl.gov/cmip5/) to learn how to do this.
- e) I acknowledge the potential limitations of the data obtained from this archive. These may include (but are not necessarily limited to) errors in the models, shortcomings in the experiment designs, the conjectural quality of the forcing scenarios used to drive the models, and so on.
- f) I understand that although the model output has been subjected to a quality control procedure, unrecognized errors almost certainly remain.
- g) To aid participating groups in understanding and improving upon their models' behaviors, I will respond to reasonable requests from the WGCM for feedback about my CMIP5 research results (e.g., reporting model deficiencies, recording CMIP5 publications, etc.).
- h) Although I may freely share downloaded CMIP5 data with close collaborators, I understand that I may not redistribute the data more widely without abiding by additional terms of use enumerated below.

Users may share data with close collaborators who have agreed to abide by the above terms of use. A research institution wishing to share CMIP5 data *internally* among its staff may seek permission to do so from PCMDI by submitting to PCMDI this form permission_to_share_form.docx. Others planning to redistribute CMIP5 model output must abide by additional "terms of use" enumerated below.

***Additional* terms of use for redistribution of CMIP5 model output:**

For the following reasons, users are discouraged from downloading CMIP5 data for the purpose of redistributing it to others (beyond their close collaborators):

- The CMIP5 data archive is a dynamic collection of files, and it will be difficult to keep a copy (of even a small subset of the archived data) up to date. (Even if an automated update procedure is implemented, the resources allocated to the official CMIP5 archive could be unduly affected.)
- The modeling groups have requested that users downloading CMIP5 data be registered and agree to the terms of use, and PCMDI alone is responsible for this.

If despite the above arguments against it, you wish to redistribute CMIP5 data to others, there are specific conditions that must be met. In addition to abiding by the terms of use (see above), anyone redistributing CMIP5 output (beyond their close collaborators) is required to:

1. Seek permission to proceed by contacting PCMDI (taylor13@llnl.gov, williams13@llnl.gov).
2. Record contact information for all users downloading CMIP5 output. These records must be sent to PCMDI quarterly in a format acceptable to PCMDI. These records will be used 1) to inform users when flaws in model output are discovered, and 2) to gauge the impact of CMIP5 results through the collection of usage statistics, as requested by the modeling groups.

3. Continually update data holdings to accurately reflect the CMIP5 archive. This will prevent known flawed data from being distributed.
4. Require users to agree to the "terms of use" and "acknowledgement" statements found at: http://cmip-pcmdi.llnl.gov/cmip5/terms.html .
5. Display a prominent banner showing the source of the data (CMIP5) and indicating that the original CMIP5 data can be accessed through the ESGF data portals (see http://pcmdi-cmip.llnl.gov/cmip5/availability.html).
6. Display a warning that the modeling groups have not checked or approved of the data being distributed.

In general, the operation of ESGF and its performance in serving other users must not be adversely affected.

Note that commercial use of all but "unrestricted" data is strictly forbidden. In particular, paid advertisements must not appear on websites providing access to data (or derived data products) from CMIP5 models that have been designated as being available only for "non-commercial research and educational purposes".