EUROPEAN COMMISSION
Community research

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

IS-ENES2 DELIVERABLE (D -N°: 5.3)

Report on basic data access protocols and data quality control

File name: {IS-ENES2_D5_3.pdf}

Author: ***F. Toussaint***

Reviewers: **B.N. Lawrence**
Christian Pagé

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

Date for review: ***19/10/2016***

Final date of issue: ***25/11/2016***

Revision table			
Version	**Date**	**Name**	**Comments**
0.1	2016-08-17	Frank Toussaint	First formal release
1.0	2016-11-21	Frank Toussaint	Entered hints of reviewers Lawrence/Pagé

Abstract

The deliverable aims at summarizing the present status of standards in various projects of Earth System science. It refers to data quality and data access, legal and technical, in order to ease interdisciplinary data exchange from/to the European climate community. The situation is refereed with respect to various European projects. Different aspects of existing standards are reviewed, for a legal standard a recommendation is given. This report is based on the results of interviews and literature analyses as given in Milestones 5.2 on *Consultation on Data Access Protocols* and 5.3 *Consultation on Quality Control Requirements* of the IS-ENES2 project.

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

For any data interchange, standards are necessary. Such standards are especially necessary for the field of climate model data as data are re-used in many parts of the society. In the following we will discuss legal, technical, and quality standards.

As standardized legal conditions of use (ToU) the general use of Creative Commons' CC BY license is recommended, as a wide spread, well supported, easy to handle, and very open license text. Where inevitable, the non-commercial form CC BY-NC may be used[1].

Besides legal standards, technical standards for climate model data are needed. They are set by different sources; they have evolved since several decades. For climate model data dissemination, the Earth System Grid Federation (ESGF) drives many parts of the necessary standardisations on an international framework level. The introduction of White Papers from the WGCM Infrastructure Panel (WIP) lead to a strong improvement of the detailed specifications from CMIP Phase 5 to Phase 6 – in quantity, quality, and robustness. The conformance requirement should be further strengthened.

Finally, for comparison of the data, quality standards are desirable. Here various different aspects of quality are to be distinguished: the scientific quality of the data themselves, quality of the different types of metadata, quality and adequateness of the data format, easiness of data access, and much more. For all of them yields: standards for them are difficult to set. So evaluation and description of quality measures in the metadata is important. The final decision on the data's usability will remain with the data user as it depends on his/her objective. A present approach to describe quality is the use of a Quality Maturity Matrix which shows the values of different dimensions quality has.

[1] This in single cases can be a strong constraint (see chapter 4).

2. Introduction

In a world with increasing importance of data and information, the exchange of these data between the scientific communities becomes more and more important. There are, however, only a limited number of widely agreed legal, technical, and quality standards. Instead, many projects have evaluated and assessed existing standards before they developed their own. So interlinking between the data and metadata (MD) of different research partners often is still difficult.

To enhance mutual understanding and interaction between different projects and institutes, researchers and data centres decided for a deliverable in the frame of IS-ENES2 to enhance the exchange of information with other parts of the community. IS-ENES2 WP5/NA4 tries a twofold approach to these problems.

Firstly, there are the general efforts between the IS-ENES data centres and the data producers on homogenisation of standards which include legal aspects as well as technical standards for automated data access and data handling. Here the ESGF data dissemination system has put a de-facto standard on the technical level. In addition, on governance level there are strong tendencies to a better homogenisation of Intellectual Property Rights (IPR), e.g., in CMIP by the WIP[2].

Secondly, many other projects already dealt with legal and technical standards. Here we give an overview of some of their products and the relation to possible issues in all three fields: legal, technical, and quality. They have to be discussed and related to ENES' practical work. For the discussion of IPR this not only refers to the development of own project specific guidelines like in CMIP/CORDEX. There also exist some comparison projects like COOPEUS.

We want to stress, that a comprehensive, fully consistent system of legal and technical standards as well as their control can be an important topic for the newly founded Data Task Force of IS-ENES2 and should be pursued there, too.

[2] WGCM Infrastructure Panel (WIP), B. N. Lawrence 2015: White paper on *CMIP licensing and Access Control*, https://www.earthsystemcog.org/projects/wip/resources/ – Papers – Final Versions [on 2016-08-04]

3. General Considerations on Legal Data Access Agreements

In the past years, there has been a tendency in the public opinion towards more openness for data which are gained with public money and for scientific data in general. In addition, the balance between property rights on one hand and visibility of the author on the other has been discussed and scientific merits grow with visibility of the work and its author. This led to a further drive to more open data. Especially, as many funders used their strong influence on the data access policy of a project to urge the project partners to open their products for re-use.

There are, however, still many details to clarify – particularly in scientific fields in which data transfer between different research branches (and even with administration, politics, and the media) is common. So shared legal standards on rights to copy, to use, and to adapt intellectual works get more and more important.

This chapter starts with a view on IPR[3] in other projects and discusses some related topics. The following chapter will set out a proposal for IPR for climate model data in Model Intercomparison Projects (MIP) and others.

3.1 The View from Outside: Data Access Protocols of Global Organisations and Projects

For data interchange, commitments of international organisations and projects are manifold. Here we refer to some of those statements that are of special relevance for IS-ENES and climate science.

There is, however, one important aspect of IPR principles that should be mentioned here: Only a minority of them refers to users and user groups with respect to their rights of usage. The majority refers to the use itself directly, mainly distinguishing between commercial and non-commercial use. They are mostly independent of the individual that undertakes the use but refer to the use of the data. In this sense, the author's right to influence the form of use is higher rated than a possible right to discriminate between users.

However, to differentiate between (non-)commercial use and (non-)commercial research is not trivial and sometimes might unwillingly (?) exclude users which do not work on public money, as, e.g., some Non-Governmental Organisations (NGOs) or someone who is paid a public salary but working on a subcontract which might be partly commercially funded. They sometimes will need to sell their products by getting at least compensation for their own personnel costs which is already regarded to be a form of commercial use. This principle is very explicit in the rules of the Creative Commons (CC), whereas it is much less followed in some projects.

[3] Intellectual Property Rights

3.1.1 The global view: General statements of legal principles from global organisations

In 2003, the **Berlin Declaration** on Open Access (OA) to Knowledge in the Sciences and Humanities (22 October 2003) was signed by first partners then by many other governments, universities, research institutions, funding agencies, foundations, libraries, museums and archives. It yields for data as well as for written publications and is one of the milestones of the Open Access movement. Today more than half a thousand signatories commit themselves to the *free, irrevocable, worldwide right to access, copy, use, distribute, transmit and display the data for any purpose* but with proper attribution.

There is a general tendency to make scientific data more open. So the policy of the **International Council for Science – World Data System (ICSU-WDS)** is *full and open exchange of data, metadata and products… with minimum time delay and at minimum cost*.[4]

The **OECD Guidelines** contain another interesting aspect. In addition to general requirements like openness, IPR, interoperability and quality, they give a list of recommended or at least acceptable limits of data access for research:

- National security: intelligence, military, political
- Privacy: data on human subjects
- Trade secrets including IPR: confidential data in business and other
- Protection of endangered species: location data (for protection sake)
- Legal process: data under consideration in legal actions.

However, within the scope of ENES the OECD limitations assumedly will not play a role; for other Earth System Sciences the protection issue may be relevant.

The Position Statement of the American Geophysical Union[5] claims that *Earth and space science data should be widely accessible in multiple formats and long-term preservation of data is an integral responsibility of scientists and sponsoring institutions*. They give a wide interpretation of the word data referring to the data described in their journals. It includes data derived from third sources as well as the software which is producing the data.

3.1.2 EU's Horizon 2020 requirements

As one of the main funders of European projects, the EU has strong influence on any IPR agreement approved by European projects. The general EU guideline "as open as possible and as restricted as necessary" is explicitly pointed out in various places. [6]
Some more detailed rules are given as[7]:

[4] ICSU World Data System, WDS Scientific Committee (2015-11): WDS Data Sharing Principles, DOI: 10.5281/zenodo.34354
[5] http://sciencepolicy.agu.org/files/2013/07/AGU-Data-Position-Statement_March-2012.pdf
[6] https://www.iprhelpdesk.eu/faq – Horizon 2020 – Access Rights [2016-08-02]
[7] EUROPEAN COMMISSION, Directorate-General for Research & Innovation, 2016-07-26: *Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020*

Access rights to background and results for the implementation of the project shall be given to other beneficiaries until the end of the project, even by the participants that leave the project before its completion.
On the other hand, requests for access rights to background and results from other beneficiaries for exploiting their own results shall be made up to one year after the end of the project.
Exceptions are where these requests are in collision with superior rights like duties of secrecy (see 3.1.1). However, the latter usually will not apply for Earth System model data.

For open access to (article) information the EU distinguishes between two types of OA:

> ***Green Open Access:*** *Self-archiving (also called 'Green' open access) means that the published article or the final peer-reviewed manuscript is archived by the researcher – or a representative - in an online repository before, after or alongside its publication. Access to the article is often – but not necessarily - delayed ('embargo period') as some scientific publishers may wish to recoup their investment by selling subscriptions and charging pay-per-download view fees during an exclusivity period.*
>
> ***Gold Open Access:*** *Open access publishing (also called 'Gold' open access) means that an article is immediately provided in open access mode by the scientific publisher. The associated costs are shifted away from readers, and instead to the institute to which the researcher is affiliated, or to the funding agency supporting the research.*[8]

3.1.3 Results of the comparison of EU/US legal standards in COOPEUS

The European and US project COOPEUS compared the legal standards of about a dozen institutes and research infrastructures (RI) in Europe and in the United States. On their findings a deliverable was produced[9]. In a second deliverable[10] this is analysed and a *Joint core data and IPR policy* is manifested. As an orientation, reference is given to the statements of the Global Earth Observation System of Systems (GEOSS).
In addition to free and open access, data for free, and as soon as possible, the resulting COOPEUS policies comprise publicly available MD in international accepted standards and attribution for resources. IPR and international legal and ethical frameworks are respected; including the commitment to indicate them clearly with the corresponding MD.
In addition, COOPEUS gives a template for a Memorandum of Understanding[11] (MoU) for the partner organisations, referring to forms of research, operating procedures, implementing arrangements and exchange of information.

3.1.4 Legal standards in CMIP5

In the IPR for CMIP5 declare a subset (about three-quarters[12] of the models) for unrestricted use[13], provided appropriate attribution. A detailed description can be found in B. N. Lawrence[14] .

[8] Fact sheet: Open Access in Horizon 2020, European Commission 2013-12-09 [2016-08-02]
[9] COOPEUS, Deliverable 7.1, www.coopeus.eu [2015-05-25]
[10] COOPEUS, Deliverable 7.2, www.coopeus.eu [2015-05-25]
[11] COOPEUS, Deliverable 7.4, www.coopeus.eu [2015-05-25]
[12] This figure has increased during the project from a start value of about 1/2.

The data of the remaining institutes are restricted to non-commercial research and educational purposes. For the latter, selling of material is not permitted.
Aspects of reproduction costs and NGOs etc. were not taken into account and might in given cases need interpretation.

3.1.5 Legal standards in CORDEX

As both projects are closely related, the Terms of Use/IPR for the Coordinated Regional Climate Downscaling Experiment[15] (CORDEX[16]) were mainly designed following the standards of CMIP5. So here the same comments apply.

3.1.6 IPR in the CHARMe project: the Open Annotation

A special case is the CHARMe[17] project. Here the data is annotations to scientific data, made by users and/or data providers. The Conditions of Use[18] state that the author of an annotation retains the copyright of his/her annotation. If there is no further commenting on this, it may be doubtful for third parties, whether or under which conditions they may cite the annotations. The initial aim here was, to make clear, that CHARMe does not claim any of these content related rights.

3.1.7 Present situation of legal standards in EUDAT

In the European Data project (EUDAT) the role of data depends on the type of service which is offered by EUDAT. Accordingly, the legal issues depend on the service, too. The following three examples of EUDAT services might illustrate this.

- Be to find (B2FIND)
  For this big catalogue the MD is harvested at the sites of the data providers. All MD are completely open. This is fixed as a verbal agreement at the time of first contact between the data centre and the MD provider. Other MD would not be welcome. Unlike the MD, the data are not necessarily open. However, the B2FIND catalogue only links to the data providers. So within EUDAT there is no necessity to handle IPR of the data themselves.
- Be to share (B2SHARE)
  This service offers data upload for low volume data (long tail data). The clients have to upload their data themselves, knowing that this is a portal for dissemination of open data. During the upload process they can select from a list of various standardised open licenses or can upload an own license text – which not necessarily is completely open.
  In case no further information is given by the uploader, the act of uploading is regarded as a conduct implying intent[19] – the intent to open the data under the indicated license.
- Be to save (B2SAVE)
  The storage of high and medium volume data which is offered by this EUDAT service

[13] CMIP5 Terms of Use, http://cmip-pcmdi.llnl.gov/cmip5/terms.html
[14] B. N. Lawrence, WIP White Paper (as above footnote 2)
[15] CORDEX Terms of Use on https://madwiki.dkrz.de/CORDEXDataManagement
[16] CORDEX: http://wcrp-cordex.ipsl.jussieu.fr/ , data page: http://cordex.dmi.dk/
[17] http://www.charme.org.uk/
[18] https://charme.cems.rl.ac.uk/conditionsofuse/ [2016-08-04]
[19] In German law: *konkludentes Handeln*

requires some care for IPR. However, as here always is an interaction between repository and data provider, their direct contact enables both to find a common solution for this in case the data are not planned to be completely open. This form to handle IPR obviously has the disadvantage that the data repositories might have to deal with many different forms of licenses for their customers.

3.2 Different rights for different communities – what about commercial use?

Data are widely interchanged. This is especially valid for the Earth System data of climate projections as they are used in the fields of education, administration, politics, industry, and of course many other parts of Earth Science. Nobody can want different rights of re-use for different communities.

This calls for a licence which is wide spread in society and free of special geoscientific concepts. Especially, in practice one will not be able to enforce rules which refer to, e.g., the differences between dynamical and statistical downscaling.

A special case is the commercial use. This term sometimes is explained by the contrary which is defined as: *Results from non-commercial research are expected to be made generally available through open publication and must not be considered proprietary*.[20] By this definition one tries to hinder unwanted exploitation of the data product. On the other hand, the exploitation of one's self-made data products is widely accepted. In projects it is granted, e.g. by EU to its beneficiaries and political promotion of research is often also promotion of economics by spin-off. However, some research corporations are put by their funders in the contractual position that they are forced to a maximum exploitation of their results[21]. This implies in their view that at least for not externally funded projects they only can disclose the results for non-commercial use. In these cases it seems to be fair to impede the selling of the data and their products but to allow other forms of re-use. This should be independent of the licensee; instead, it should depend on whether or not there is an exploitation of the data by a third party which includes some profit.

The Creative Commons define *non-commercial* as *not primarily intended for or directed towards commercial advantage or monetary compensation*.[22] In case of a share-alike license[23], for derivatives this obviously excludes a monetary compensation for labour costs if not even for material. So here the non-commercial element in the license will make it difficult for some NGOs and other non-profit bodies to draw derivatives, whereas this is not a problem for cases where money is not an issue. Thus industrial but also governmental bodies are preferred.

[20] CMIP5 Terms of Use, see Chapter 3.1.4
[21] See German law of the DWD: DWD-Gesetz, §6(2), http://www.gesetze-im-internet.de/dwdg/__6.html
[22] E.g. https://creativecommons.org/licenses/by-nc/4.0/legalcode [2016-08-05]
[23] A share-alike license entitles the licensee to build own works on the licensed material. He has, however, to publish them under the identical license.

In any case, a form of licensing prohibiting the use of publicly funded data not only for commercial re-selling of the data products and their derivatives but confining it to non-commercial research work may not only be seen as iniquitous in the view of the tax-payer. It also excludes non-governmental bodies which sometimes is not intended and often is inappropriate. This may be one of the reasons why the CC BY-NC license refers to making money of the data products themselves, not to a missing non-profit characteristic of the licensee. The latter well can be a non-commercial body or a company.

Another topic which is not covered by the term *non-commercial* in most jurisdictions is the right to present works together with other things at the same time. To demand from a licensee who is allowed to show the work to third parties that he/she does not show certain other things at the same time, is a request of doubtful value. Here an example is the norm, not to show commercials on webpages that present certain research results.

3.3 Derivatives

The right to draw derivatives from a work is one of the central rights of cultural freedom[24]. However, it is difficult to give an exact definition what to call a derivative work:

> *In copyright law, a derivative work is an expressive creation that includes major copyright-protected elements of an original, previously created first work (the underlying work). The derivative work becomes a second, separate work independent in form from the first. The transformation, modification or adaptation of the work must be substantial and bear its author's personality to be original and thus protected by copyright. Translations, cinematic adaptations and musical arrangements are common types of derivative works.*[25]

Here at least the words *major* and *substantial* need more detailed definitions depending on the situation. A picture of the surface pressure at a certain time step, drawn from a 100 years' model run is certainly not a major part of the latter and so is not a derivative. A movie of the full 100 years, however, builds on a major part of the underlying work and can well be substantial – depending on the effort made. An artist's view certainly bears its author's personality.
Similar problems occur when we do not just discuss the quantity of the underlying work but the quality. A complete visualisation of a Global Climate Model's run is a derived work. But what about using it to force Regional Climate Models (RCM)? The RCM certainly is an own work and so is its output. But this view sometimes seems to be in the eye of the beholder. And who is to control what a licensee will do with the downloaded data? So where to draw the limits?

[24] http://freedomdefined.org/Definition [2016-08-04]
[25] Wikipedia at: https://en.wikipedia.org/wiki/Derivative_work

This shows, that putting restrictions on derived data leads inevitably to a certain amount of legal uncertainty and thus to a wide field of work for lawyers. This is why this paper strongly recommends not restricting any derived uses of the data. The understandable desire to prevent improper use by layman or intended abuse by others should not lead to a general prohibition to all re-users in whatever aspect.

3.4 Enforceability of standards and Terms of Use

Compared to the number of cases in which data are used that have an underlying license regulation, the number of court cases and judicial proceedings is remarkably close to zero. Unlike the situation in other fields of life, in science this is obviously due to the fact, that there is a high evolved standard for citation and attribution whereas at the same time authors have a strong interest to be cited and thus have their data used.

For the effectiveness of Terms of Use (ToU), it is important that the *prospective users are most likely to have seen, or know of, the appropriate license under which they use that digital property*[26]. Here besides clear statements in the download path one can include the ToU into the data objects themselves as proposed by B. N. Lawrence in the respective paper of the WIP. However, in case the rules are only referenced and not cited in the file (e.g. by URL), new problems might occur. Perhaps the most important is, that a data distributor cannot tell its customers anymore, what really is behind that reference, if the referenced norm is not a common standard. Any special features in the rules of the data producer stay in his sphere and mostly cannot be commented or judged by the data disseminator. Furthermore, any changes of those rules after publication are beyond the data centres control. All this is a strong reason to decide for a common wide spread standard, which, of course, will be versioned.

A strong reason to not prosecute those cases is the imponderability that comes with unclear definitions. This is especially true for non-standard regulations that are rarely interpreted by courts (see above) but to a lower degree also for well-known standards.

In addition, as these cases of license law usually do not touch questions of criminal law, the civil law lets the presumably damaged party on its own prosecution means.

The above said leads to the advice to make the pursuing of license rights up to data producer, i.e. to the license author. Any lack of clarity or availability of the license rules will be in the responsibility of the creator. This is especially useful when a data producer does not stick to common license standards but includes or adds some deviations or even has a home-knit version. There is no reason to make those judicial issues up to downstream data centres.

[26] B. N. Lawrence, WIP White Paper (as above footnote 2)

4. Proposal Regarding Standardized Conditions of Use

4.1 Summary of demands

Summing up which are the most important demands that are advisable in connection to the implementation of the standard license text (ToU) to be proposed, we come to the following recommended list of requirements, ordered by content, form, and application of the norm.

1. Content related criteria

a) The license text must include a disclaimer.
b) The license text must include a citation requirement.
c) The license needs to include rules for databases and collections.
d) The license should not discriminate against non-profit organisations.
e) The license should not impede the derived uses of the data.
f) The license should take into account the wide spread use of climate model data outside science in administration, politics, and necessary public discussions on climate change.

2. Formal criteria

a) The license text needs to be maintained by some corporate body, as the world around it changes over years.
b) There should be user support available for help and advice in applying those rules and to explain them in case of any doubts.
c) For three reasons the license must have good spread in society:
   - The rules should be compatible to what is used outside the community like in other research institutes, administrations, and industry,
   - for legal certainty, a consolidated jurisdiction is required,
   - long term reliability of the services of maintenance and user support is more probable for wide spread standards.
d) The license should be as easy as possible to handle for its users.
e) The license needs to work globally.

3. Application

a) As every data user should realise the Terms of Use (ToU), the ToU should
   - have a prominent place on the web page,
   - be included in the data objects, and in particular not be available only by reference from the data objects, unless a long term stability of the reference can be guaranteed as is the case for the big well known, stable and maintained licences.
b) An agreement on common ToU in a project is advisable as, e.g., it fosters user support by the data centres.

4.2 The Creative Commons license

A group of free and easy-to-use copyright licenses is offered by the Creative Commons Corporation (CC). All six offered licenses require attribution (BY), three come in a non-

commercial (NC) flavour, the others don't, and both versions (CC BY and CC BY-NC) can be supplemented by one of the two claims that derived works are not allowed (CC BY-ND or CC BY-NC-ND) or they are allowed but only when the new product is shared under the same license (CC BY-SA or CC BY-NC-SA). This paper recommends the use of CC licenses for IPR of MIPs.

The main advantages of the Creative Commons set of elements (rules) are:

- They are maintained/curated by the Creative Commons Corporation,
- they are widely used in the society, including outside science,
- CC user support is available via mail from the corporation. This, of course, should not be misunderstood as a pro bono legal service.

Which license to use? Relating this to the requirements above shows that acknowledgements are essential for climate model data, whereas the non-commercial attribute might be necessary in few cases only. Furthermore, due to the above mentioned, derivatives should be possible, as well. Two options are clearly relevant: "CC BY" and "CC BY-NC", but should they be supplemented by the *share-alike element* (SA), which claims for distribution of remixed, transformed, or built-upon work to be distributed under the same (CC…) license as the original?

This is probably no problem in the world of non-commercial research. However, climate model data of climate projections is widely spread in society outside of science. It is used by administrations, politics, journalists, and in many parts of the public discussion on climate change. Here the share-alike qualification can have far-reaching consequences. Visualisations of substantial parts of the data can only be sold on a media cost basis (CC BY-NC-SA) or non-profit bodies are forced to expose their derivatives to commercial use, too (CC BY-SA).

Here an example for the more restrictive non-commercial license (CC BY-NC-SA) might help to clarify the situation: Films that build on (e.g. visualise) climate model data cannot easily be produced by non-profit bodies. As discussed in Chapter 3.2 they perhaps were allowed to get their material cost refunded by selling the product but no labour costs. And they of course were unable to sell a DVD of those movies at market conditions. On the other hand, this would not be of any problem for industry and other big companies. Furthermore, the licensee might want to give its derivatives a maximum spread in the public – perhaps to commercial redistributors as well. Why should one hinder them?

Given the social relevance of the global discussion on climate change and on the data that supports this, the author does not recommend to issue NC data with the SA-attribute. This even more as climate science is under the critical view of the society regarding transparency and as any judicial obstacles cost society's strength in the public discussion. Instead, if a non-commercial restriction is needed, the proposal is to use just the CC BY-NC licence (as opposed to the use of CC BY-NC-SA).

For the use of the share-alike element together with only attribution (CC BY-SA) issues are less obvious. The obligation to share possible derivatives on a share-alike basis together with the right to use the data commercially means that licensees cannot issue any derived products for non-commercial use only. There might be some doubt whether this is always appropriate and intended by the licensors as it also hinders non-profit licensees to give out their derivatives on a non-commercial basis.

Hence, the proposal here is to use just the CC BY licence, which is called by Creative Commons "*the most accommodating of licenses offered. Recommended for maximum dissemination and use of licensed materials*.[27]"

Regarding standardized conditions of use (ToU) the proposal is:

1) **to generally use the CC BY license of the Creative Commons as a wide spread, well supported, easy to handle license text,**
2) **to use the non-commercial form CC BY-NC only where inevitable,**
3) **not to put any restrictions on derived works, be it by the share-alike (SA) element or by other means,**
4) **to implement the measures to ensure the perception of the ToU by the licensees (above 3. a) and b)).**

[27] https://creativecommons.org/share-your-work/licensing-types-examples/licensing-examples/

5. Technical Standards for Data Access

To ease data exchange not only common legal standards on the political level are needed. On the level of practical work the technical standards are not less important. In the following chapters we will have a look at data formats, MD formats and at exchange interfaces.

5.1 Global view of technical standards

There are various ways standards may come from. Standards by authority (ISO, government/law), standards by business power (Microsoft, Google), and de-facto standards on which the community has agreed. A special mixture of de-facto standard by power comes into play when one or more members of the community has the resources to make the use of a certain standard highly beneficial, e.g., by providing software for users of a certain data format. This is the case for the NetCDF format which is maintained by the University Corporation for Atmospheric Research (UCAR)[28] in the US which, e.g., is supported with software by UCAR and others.

5.2 The technical standards at other projects: COOPEUS, CMIP5/CORDEX/other MIP, EUDAT

5.2.1 Results of comparison of EU/US technical standards in COOPEUS

The European and US project COOPEUS compared the technical standards of about a dozen institutes and research infrastructures (RIs) in Europe and in the United States. On their findings a deliverable was produced[29] which includes an *Interoperability Maturity Index*. To evaluate this for a given institute/Research Infrastructure (RI) a set of 15 questions is given, including the appropriate form on the website.

The comparison of a set of 12 institutes/RIs led to the following results for compliance to MD standards (including planned):

MD representation:	pure ASCII 8/12,	xml 6/12,	JSON 3/12
MD format:	Dublin Core 4/12,	ISO 19115 5/12,	DIF 1/12
MD interface	OAI-PMH 2/12,	CS-W 2/12,	Open Search 5/12

The analogous comparison for data standards led to the following results (including planned):
Data formats: Pure ASCII 10/12, NetCDF 8/12, HDF5 6/12, SEED/miniSEED 3/12
For details and further comparisons see the deliverable.
For data, COOPEUS concludes that NetCDF has gained much importance in the COOPEUS community and has good potential to become the de facto standard of the project.

5.2.2 Technical standards in CMIP and CORDEX

The climate data of these projects (CMIP5 and the present CORDEX) are disseminated by the Earth System Grid Federation[30]. Here the sophisticated system including the THREDDS Data

[28] This situation is comparable to that since some decades in the astronomical community with respect to NASA's FITS data format and in the field of software to MIDAS (maintained by the European Southern Observatory, ESO).
[29] COOPEUS, Deliverable 7.3, www.coopeus.eu [2015-05-25]
[30] see ESGF, https://www.earthsystemgrid.org

Server[31] which is also supported by UCAR puts demands on format and structure of the data. To check them, is an important part of the data ingestion (see IS-ENES2 milestone 5.3).

There is a variety of technical standards to be explicitly defined in projects that want to publish their data in the ESGF system (see the definitions on the CMIP5[32] website and the CORDEX Data Management Specifications[33]).

Data Reference Syntax

The Data Reference Syntax (DRS)[34] is an hierarchy of sets of keywords specifying the single dataset. Keyword sets describe, e.g., the project, institute, model, ensemble member, variable etc.

Data format

Although THREDDS is able to work with WMO-GRIB format, too, CMIP uses NetCDF for all data. The MD in the file headers has been normalised. However, in CMIP5 sometimes data providers did not comply with these instructions. For CMIP6 better compliance probably will lead to more robust workflows.

Metadata format

The primary metadata format within ESGF is provided from information held in the files.
As this information is harvested by the THREDDS and put into a database, its output can be formatted as necessary. There are at least (users can add their own extra metadata) three levels of ESGF metadata required: (1) ESGF requires the use of the climate model output rewriter, CMOR[35], to include important data information, as well as (2) the DRS information, and (3) the CF conventions[36], an extension to the NetCDF format. The conventions for CF (Climate and Forecast) metadata are designed to promote the processing and sharing of files created with the NetCDF API and are increasingly gaining acceptance.
In addition, the ESIP federation[37] specified necessary attributes in the NetCDF file headers in the Attribute Convention for Data Discovery[38] (ACDD).

5.2.3 Technical standards in EUDAT

Some of the standards used by EUDAT are wide spread and accepted. However, here different standards have to be distinguished as EUDAT comprises scientific fields of humanities, natural sciences, and medical sciences.
For metadata harvest the well-known protocol OAI-PMH[39] of the Open Archives Initiative is used. XML formatted metadata of possibly proprietary structure are mapped to the EUDAT catalogue standard. The metadata then are SOLR-indexed and can be searched by the users (B2FIND).
In B2FIND, B2SHARE, and B2SAVE the data are kept in the formats of the data provider. Here no further standards are applied.

[31] TDS, https://www.unidata.ucar.edu/software/thredds/current/tds/
[32] http://cmip-pcmdi.llnl.gov/cmip5/modeling_overview.html?submenuheader=2
[33] https://madwiki.dkrz.de/farm/CORDEXDataManagement
[34] Data Reference Syntax, Taylor et al: http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf
[35] http://www2-pcmdi.llnl.gov/cmor
[36] http://cfconventions.org/
[37] www.esipfed.org
[38] See http://wiki.esipfed.org/index.php/Attribute_Convention_for_Data_Discovery_1-3
[39] See http://www.openarchives.org/OAI/openarchivesprotocol.html

5.3 Technical standards on CMIP6: A short outlook

The main difference between phase 6 of the CMIP project and the preceding phases is the collective effort of the partners to go for detailed and robust standards. Here probably the experiences of CMIP Phase 5 bore fruits as inhomogeneities in the millions of datasets in many cases caused extra manual work.
Many of the requirements are described in detail in the WIP White Papers on the project website[40], so e.g., Global Variables and Controlled Vocabularies (CVs) which are central standards of the project. The framework, however, is set by the Earth System Grid Federation (ESGF[41]) as a de-facto standard. To enable the easy use of the more detailed specifications by all partners and their software tools, low-threshold access is essential, e.g. by an http accessible repository for the various lists of agreed definitions like Controlled Vocabularies (CV) of the DRS elements.

Further technical descriptions are found in detail in the different WIP White Papers.

[40] https://www.earthsystemcog.org/projects/wip/resources/
[41] http://esgf.llnl.gov/

6. Standardized Quality Assurance

In a world of exploding amounts of data the quality of these data becomes more and more important. There is, however, no common sense of what quality is and how to measure and compare it between different research partners.
The three main reasons for improving quality standards are the strong increase of data quantities over the last years, the multiple data reuse, and the desire to review scientific works after a couple of years.
There are general efforts between IS-ENES and partners on quality control and quality control requirements which include technical requirements for automated data access and correct data indexing by identifiers (DOI). In addition, on governance level there are strong tendencies to better data homogenisation, e.g., in CMIP by the WGCM Infrastructure Panel[42].

6.1 What is data quality?

One has to keep in mind that the quality of data strongly depends on the intended use of them. When the data description standard of ISO 19157 already states:

> *This International Standard recognizes that a data producer and a data user may view data quality from different perspectives. Conformance quality levels can be set using the data producer's product specification or a data user's data quality requirements*

one might want to add that even different users see the data quality from different perspectives. The judgement depends on different intended purposes: user1 vs user2 and both of them vs the data producer. This is where the detailed description of data quality aspects becomes essential to make the data fit for reuse: The proof of the data quality is at the user.

However, one can list various aspects of quality that are important for some or all users. This firstly can be split into

- General Aspects like adequateness of the format and the structure of the data, technical accessibility, adequateness of the coordinate system used,
- Data Aspects like accuracy, completeness, possible errors, error bars, conformance to measurement requirements,
- Metadata Aspects like completeness (richness), versioning, consistency of MD, conformance to MD standards.

To disentangle these different aspects and the users' views on them is a vast field of work which goes far beyond the scope of this paper. The complexity of this is perhaps the reason, why the analyses of COOPEUS did not result in very detailed answers on questions related to this (see below 6.3.2).
And perhaps finally one has to rely on the user community to feedback what is good and bad in the view of the respective user.

[42] WGCM Infrastructure Panel, F. Toussaint: White paper on *CMIP Quality Assurance*

6.2 Quality assurance as part of the workflow

Quality Assurance needs quality controls all along the workflow and in all parts of it. Here Quality Assurance (QA) for data differs from QA for metadata (MD). For the data values its creator (author/editor) has the main responsibility for correctness and scientific quality. Here the concept of quality depends on the use of the data – so it has to be related to the accompanying MD which contain information on what the data is adequate for. As MD is concerned, the data centres partly have the responsibility for QA. They should keep track of, e.g., completeness and comprehensibility. Within a project the agreement on these responsibilities should be part of the data management planning.
In a project it is essential for the planning of data management to develop and publish quality assurance criteria and record their results. This is for MD even more important than for data.

6.3 Quality aspects in scientific projects: COOPEUS, CMIP, CORDEX, EUDAT & KomFor

To find out about the general present situation of quality control and quality control requirements in neighbouring projects, for CMIP, COOPEUS, EUDAT, CORDEX and KomFor interviews personal and by phone were conducted and documents reviewed. They are summarized in the following.

6.3.1 Situation of quality control requirements in COOPEUS

The project COOPEUS[43] (Strengthening the cooperation between the US and the EU in the field of environmental research infrastructures) which ended in August 2015 was a transatlantic partnership of infrastructure systems funded by EU and NSF. It covers some of the topics here in question. The project provides an analysis of partner policies which often focus on IPR. Most of the outcome of the COOPEUS project is presented on the website as project deliverables.

As far as standards of data quality are concerned, COOPEUS gets in its Summary Report[44] to the conclusion that there is *need of common quality control and assurance plan. This task is addressed partly within COOPEUS, but will need longer period collaboration*. The underlying evaluation of the survey of different Earth Science institutes[45] states that most of the questioned institutes have at least partly quality assurance measures implemented in their workflows[46]. Here some investigations that go further than COOPEUS did would be beneficial. This refers to the definitions used for *quality* and comparison of measures and checks.

[43] See www.coopeus.eu
[44] COOPEUS, Deliverable 8.2, www.coopeus.eu [2015-05-25]
[45] COOPEUS, Deliverable 3.1, www.coopeus.eu [2016-08-03]
[46] COOPEUS, Deliverable 8.2 p5, www.coopeus.eu [2015-05-25]: *This task is addressed partly within COOPEUS, but will need longer period collaboration.*

6.3.2 Present Situation of quality control requirements in CMIP

The *Coupled Model Intercomparison Project* (CMIP[47]) compares global climate model data to improve Global Climate Models. In CMIP5, the quality of data was mainly due to the data producers. However, some checks were done by the data centres and communicated to the producers and consumers.

There was a wide range of metadata in CMIP5. Some of them, like the description of the data production environment and the producing model were filled in a questionnaire by the data creators and were mainly not controlled. However, in between a control initiative on these data has started.
Other metadata had to be mentioned in the file headers or coded in the file name and had to correspond with the agreed specifications[48] which were laid down in a spreadsheet. This template was mainly handled by an office application and was not error-free.

The checks were split into three levels. Their results were documented and stored.

In the next of the CMIP projects, CMIP6, some of these problems will be overcome by strengthening the position of the data nodes which accept or refuse the data from the producers. The plan is to entitle them to refuse data that obviously do not follow the agreed standards. The data producers on the other hand should and can get the checking software to ensure beforehand the compliance of the data to the most important ESGF standards.

Quality of archives was not an issue of CMIP5 nor is it planned for CMIP6. The quality of the dissemination system (data nodes), however, needs to and will be improved for CMIP6. In these contexts, quality of metadata and data node operation, the CDNOT was established. Details can be found in the WIP paper[49].

To describe the quality of data and metadata in geosciences, a very detailed ISO schema was published as ISO 19157[50]. For CMIP, the CHARMe project came to the recommendation[51] to alter the CIM[52] in order to allow for ISO 19157 descriptions.

6.3.3 Quality control requirements in CORDEX

The aim of the *Coordinated Regional Climate Downscaling Experiment* (CORDEX) is to foster scientific work on regional climate projections by coordinating the scientific work in this field. Like in CMIP5, the quality of data was mainly due to the data producers. In CORDEX as well, only some checks were done by the data centres and communicated to the producers and consumers.

[47] e.g.: http://cmip-pcmdi.llnl.gov/
[48] K. Taylor, see http://cmip-pcmdi.llnl.gov/cmip5/documents.html
[49] CDNOT Terms of Reference & above footnote 37, https://www.earthsystemcog.org/projects/wip/resources/ [2016-08-03]
[50] See https://wiki.earthdata.nasa.gov/display/NASAISO/ISO+19157
[51] Deliverable 400.2, http://ensembles-eu.metoffice.com/charme/deliverables.html [2016-08-01]
[52] Common Information Model of metadata

Similar to CMIP5, part of the metadata was collected from the data producers – in case of CORDEX via an Excel sheet.
The use metadata were indicated in the file headers and had to correspond with the agreed specifications which, here too, were laid down in a spreadsheet. Compliance checks of the data were conducted according to the procedure in CMIP as described in Chapter 6.3.2. In case of CORDEX the data node administrators had agreed to a special policy in a Memorandum of Understanding: In the ESGF data dissemination system no data should be published which are not in agreement with the CORDEX standards given in the agreed table.
Quality of archives was not an issue of CORDEX nor is it planned.

6.3.4 Present Situation of quality control requirements in EUDAT

The project *European Data Infrastructure* (EUDAT[53]) spans over different scientific fields. The data are disseminated via the *B2FIND* portal. In the September 2014 meeting in Amsterdam one session was dedicated to metadata and their quality, the session presentations were put on the web[54].

One important result of this trans-community approach was that the data quality here mainly depends on the data producer. There was little chance to judge on and reject data of doubtful quality. For metadata the situation was different. The necessary mapping of the metadata contents from the source community to central common schemes and ontologies in most cases let weak spots appear if there were any. So they could be corrected or sorted out as invalid.

In addition, EUDAT went for judgement on the quality of data centres. Here the decision was to mainly follow the approach and rules of the Data Seal of Approval (DSA[55]).

6.4 A Quality Maturity Matrix for Quality Assessment

Quality Assurance of data plays an important role in the data publication process as well as for data re-use. However, quality control procedures and quality documentation vary greatly among scientific data. In general, every project defines its quality procedures.

In order to make the different Quality Assessments of the projects comparable, a generic Quality Assessment System has been developed in the project KomFor[56], which itself does not handle data but runs a metadata catalogue. Based on the self-assessment approach of a maturity matrix, an objective and uniform quality level system for data is derived. It consists of 5 maturity quality levels, starting with the initial level=1.
This system now gets more and more evolved. DKRZ plans to use a first pilot for the CMIP6 data.

[53] European Data Infrastructure, e.g.: www.eudat.eu
[54] https://www.eudat.eu/programme-eudat-3rd-conference [2016-08-03]
[55] www.datasealofapproval.org
[56] www.komfor.net/qa.html [2016-08-03]

7. Conclusions

For global data interchange – not just between scientists but also with administrations and industry – legal, technical, and quality standards are needed. In all these fields already many standards exist, which is a strong argument for not generating further ones. Instead one should try to use existing standards. The disadvantages of applying an existing standard that might not fit in detail the needs of every single project partner usually are smaller than the efforts to elaborate a new one which goes beyond the existing standards.

Regarding standardized legal conditions of use (ToU) it seems fair to put the least possible restrictions on the distributed data, as these are funded by public money. So the general use of Creative Commons' CC BY license is recommended, as a wide spread, well supported, easy to handle license text. Where inevitable, the non-commercial form CC BY-NC can be used, which in single cases can be a strong constraint.

Technical standards for climate model data are set by different sources. The framework which has evolved during the past years is the Earth System Grid Federation (ESGF). As a de-facto standard it drives many parts of the necessary standardisations on an international level. The more detailed specifications have strongly improved from CMIP Phase 5 to Phase 6 – in quantity, quality, and robustness.

Quality standards are difficult to set, as quality often is difficult to define. A focus can be set on the evaluation and description of quality measures in the metadata. The final judgement, however, stays with the data user as it depends on the use of the data. A present approach is the use of a Quality Maturity Matrix which shows the assessment values of the different dimensions of the quality concept.

8. Glossary

AGU	American Geophysical Union
ACDD	Attribute Convention for Data Discovery from ESIP
CDNOT	CMIP Data Node Operation Team
CC	Creative Commons, a group of licenses
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
CMIP5, CMIP6	Coupled Model Intercomparison Project, Phases 5 and 6
COOPEUS	An EU-US collaboration project in the field of Environmental RIs
CORDEX	Coordinated Regional Climate Downscaling Experiment
DIF	NASA's Directory Interchange Format for MD
DKRZ	German Climate Computing Centre, Hamburg
DOI	Digital Object Identifier
DRS	Data Reference Syntax
DSA	Data Seal of Approval
ENES	European Network for Earth System modelling
ESIP	The Federation of Earth Science Information Partners
ESO	European Southern Observatory
EU	European Union
EUDAT	The European Data project
FITS	Flexible Image Transport System, a data format
GRIB	GRIdded Binary, a WMO data format
GEOSS	Global Earth Observation System of Systems
IPR	Intellectual Property Rights
IS-ENES	The infrastructure project of ENES
ISO	International Organization for Standardization
JSON	JavaScript Object Notation, a data format
MARUM	Centre for Marine Environmental Sciences (University of Bremen)
MD	Metadata
MIDAS	Munich Image Data Analysis System
MIP	Model Intercomparison Project
NASA	The US National Aeronautics and Space Administration
NetCDF	Network Common Data Format
OAI-PMH	The Open Archives Initiative's protocol standard for MD harvesting
OECD	Organisation for Economic Co-operation and Development
QA	Quality Assurance
QC	Quality Control, Quality Check
RI	Research Infrastructure
TDS	THREDDS Data Server
THREDDS	Thematic Real-time Environmental Distributed Data Services, run by Unidata of UCAR
ToU	Terms of Use
UCAR	University Corporation for Atmospheric Research
WDS, WDC	World Data System, World Data Centre
WMO	World Meteorological Organisation
WP	Work Package

Acknowledgements

Thanks go to the interview partners of the different projects which are

Heinke Höck (DKRZ) for KomFor,
Robert Huber (University of Bremen) for COOPEUS,
Stephanie Legutke (DKRZ) for CORDEX,
Christoph Waldmann (MARUM) for COOPEUS,
Heinrich Widmann & Hannes Thiemann (DKRZ) for EUDAT,

and especially to Bryan N. Lawrence and Christian Pagé for reviewing and helpful hints.