GEDE DO Participation Request

3.10.2018

Dimitris Koureas & Peter Wittenburg

Within the framework of the DO activities we are committed to write a few papers which may shed light on the usefulness of the DO concept. Here we give a short overview:

1. The first paper is the one written by Wittenburg & Strawn which already got quite some attraction (Common Patterns in Revolutionary Infrastructures and Data: <http://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0>)
2. The second one is currently being written by some C2CAMP and GO FAIR colleagues to look at DOs from a more computer science view.
3. **The third one will be a paper with views from scientific domains how they see the usefulness of the DO concept. Here Dimitris Koureas and Peter Wittenburg take the lead to kick it off. The hope is to have a paper until Christmas.**
4. A fourth one is currently being written by Erik Schulthes (GO FAIR) and Peter Wittenburg (RDA/C2CAMP/GEDE) about the relation between the DO concept and the FAIR principles. Erik already elaborated on this in his workshop talk.

What we have noticed in our various talks in the EC is that there is a great interest in this topic and that EC colleagues see now the potential of DO-based infrastructures to realise part of the EOSC, i.e. the purpose is to get the above mentioned papers as soon as possible without compromising on the quality of the content.

This is a request to interested scientific communities to describe their potential application of the DO concept within their infrastructure plans with the intention to integrate them where possible into the third paper to be written. Examples for such applications can be found in the GEDE Share[[1]](#footnote-1) to which you all should have access[[2]](#footnote-2). We will collect the contributions, probably discuss some questions with the proposers and then see how we can best integrate the contributions into the paper to be written. The contributions must make clear statements about the use case and the gain that are expected with the integration of the DO concept. In case of questions, don't hesitate to contact Peter (+49 15141858784).

Please submit a filled in template (see below) as is indicated below and provide references were possible. You can also attach papers to your contribution email, but these should not exceed 5 pages.

Please, submit the filled-in templates until 19. October 2018

Dimitris & Peter

GEDE DO Participation Template

## Name and Institution

Tobias Weigel (DKRZ)

Reinhard Budich (MPI-Met)

## Email and Telephone Number

[weigel@dkrz.de](mailto:weigel@dkrz.de)

[reinhard.budich@mpimet.mpg.de](mailto:reinhard.budich@mpimet.mpg.de)

## Names of Collaborators

(if applicable)

## Research Community

ENES – European Network for Earth System Modelling

## Major Goals of the Ongoing Infrastructure work

The ENES[[3]](#footnote-3) community develops and operates e-infrastructures and end-user tools and services which are used by the European but also the global climate data community. An integral part of this is the global Earth System Grid Federation (ESGF[[4]](#footnote-4)), in which ENES participates with operative nodes and contributes key developments. The ENES/ESGF infrastructure facilitates data distribution and services towards end-users for scientific community projects, including the CMIP6 coupled model experiments, whose outputs also contribute to the evaluations given in the IPCC assessment reports.

## Potential of the DO Concept for your Work

The ENES community faces manifold challenges for data management, processing, and overall aspects of making data FAIR, based on technologies from ESGF and other initiatives such as EUDAT and EOSC:

* Data volumes, but also the number of data objects, are increasing exponentially. The CMIP6 data aggregation is expected to hold more than 100 PB of data, distributed among multiple 100’s of millions of files. The key operational processes dealing with data objects must scale up, yet they rely on technologies that are difficult to automate individually. Resources will not increase as much as data object number and volume, and therefore, approaches for **automated digital object management** are highly desired. If established, automated DO management can make ENES data management much easier, more scalable, and open up further service opportunities, for instance, for automated data replication, data quality control/verification and provenance tracking.
* Data workflows are becoming more complex, covering the entire range of individual stages from data production, dissemination (e.g., via the ESGF), quality control and long-term archival, reuse and repurposing by a wide area of actors. However, there are not yet conceptually mature and operationally feasible approaches for **enabling automated workflows** and ensuring that **provenance is gathered automatically and exposed in a user-friendly way to data consumers, producers and funders.**
* Data consumers will work at **higher levels of abstraction** in the areas of data processing and analytics. ENES is reacting by offering more sophisticated Virtual Research Environments (VREs) to an increasingly diverse user base with less expertise in core modelling areas. **Tailored services** with shallower learning curves, strong user integration and good usability are needed. ENES looks towards the Digital Object concept – making DOs primary citizen – also as a means to achieve abstraction, truly separating concerns of data storage, access and management from user’s active work environments. Achieving this will ultimately help users to work more effectively without having to deal with ‘data mangling’ tasks.
* By establishing **actionable digital object collections**, the ENES data infrastructure can evolve its services. In addition, new user services can be established. For example, the ESGF already supports PID-based digital object collection in the form of **‘shopping baskets for data’**, which enable any user to clearly reference arbitrary aggregations of data they downloaded and worked with. This can, however, be only one step towards tailored services that enable users to work not only on single objects, but on collections and tightly associated graphs of objects.
* If there were agreement on the **central interfaces** for managing digital objects, their associated kernel information (key-metadata) and types, and an **offering of sustainable type registries**, much of the post-production workflows involved in producing high-quality climate data products could be simplified. For example, climate models could be enhanced so they can describe themselves; yet, without sustainable registries that store the types of models and the variables that are used in data, such effort is too difficult to maintain across multiple scientific community projects.

The ENES community looks towards initiatives such as GEDE and C2CAMP to address these points, starting with individual prototypes and small-scale solutions. There is however now a good opportunity to reach out to other communities and work towards a larger, overarching architecture and operationally capable solutions, based on the concepts and models the GEDE participants have discussed and matured in the past. Central design principles are adherence to agreed interfaces, for example based on specifications from RDA, IETF or W3C; finding the balance between complexity and feasibility, particularly regarding metadata and semantics; and building long-term sustainable services for data processing and provenance tracking.

1. <https://datashare.mpcdf.mpg.de/index.php/apps/files?dir=/GEDE/digital%20objects/GEDE-DO%20meetings/workshop-september-18&fileid=63958403> [↑](#footnote-ref-1)
2. Those who presented their plans already at the recent workshop do not have to submit these templates since we discussed the slides already. [↑](#footnote-ref-2)
3. European Network for Earth System Modelling, <https://www.enes.org> [↑](#footnote-ref-3)
4. <https://esgf.llnl.gov> [↑](#footnote-ref-4)