Supporting research data management via compute services

1 PROBLEM

Many researchers are unaware of data management requirements and services available to help meet these. Opportunities should therefore be taken to signpost relevant information when researchers are contacting infrastructure/support services. SURF is doing this via its compute calls, raising awareness of funder requirements to manage and share the associated data, and highlighting tools available from SURF and universities to help.

2 APPROACH

At this moment SURF supports about **6000 active researchers using its compute infrastructure via about 600 compute grants**.

These compute grants are part of the ongoing open call for proposals called Computing Time on National Computer Facilities. Researchers and research groups can gain access to computing time and associated data services on the national advanced computer systems provided via SURF. Researchers from all universities, university medical centres and many research institutes in the Netherlands can apply for access to this national infrastructure. Relatively small applications are handled directly via SURF, while larger applications need more extensive review via national funding organisation NWO.

Stimulating FAIR practice is conducted at various points in the application/granting process of the compute call. This is done both by provisioning information and by actively sending researchers in the compute call towards the data services of their institution.

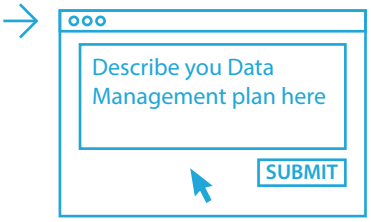
Information provisioning about FAIR practice:

a. Both SURF and NWO refer applicants (also in the application form) to research data management information (on the **website of NWO**).

b. When applying via SURF, applicants receive information about data management and which institutions use a SURF solution for storing and sharing data, including the link of the institution page. We have made an overview of institutions, **directing to the Research Drive pages of the institution**. This creates awareness for this service of the institution.

c. When applying via NWO, applicants receive several options in their application form to choose from, to use certain data storage and sharing solutions. Researchers were already making use of, for example, Research Drive or Data Archive, but until now this was just being positioned as storage attached to compute resources. We have now separated these services and **defined them as RDM services in the application form**.

d. Researchers have to **fill out a data management paragraph** (obligatory). In this step in the application process, we explain that choosing certain data services supports making their research more FAIR, such as Research Drive, Data Archive and persistent identifiers. Now, PIDs are still only assigned and used at the end of the projects.



e. When **granting periods are ending, we inform** researchers about several possibilities for keeping their data available via SURF or the institution (through automated periodic email notifications). There are, however, more options to improve FAIR information provisioning in this stage. After the grace period (about 3 months) the online access to the data is cancelled and the data is archived on tape. The idea is to work towards an automated system, in which the PI needs to either agree the data can be deleted or that the data can be transferred to an archiving infrastructure (without moving it to tape).

Sending researchers in the compute call towards the data services of their institution:

a. Compute grants are also the responsibility of the institution. Applicants receive the grant as an individual but they or their supervisor need to have a permanent contract at an institution as well. Institutions are looking for insight in who shares which data - data management has become their focus. At SURF we are **both in contact directly with the researchers via these compute grants, and at the same time we are in contact with the research support and central IT departments of institutions**. Therefore, we look for possibilities to connect these groups where this is possible. On the subject of HPC facilities, we are also in contact with other HPC computing centres in for example the USA, Germany, Sweden, Spain, Finland, Italy, etc. We see that many of these other countries also want to **bridge the gap** between offering data solutions **during the grant** (which they consider the responsibility of the HPC centre) and **after the grant** (which they consider the responsibility of the research institute or university.

b. In the internal SURF application process, we included an extra step, for each application, we check whether an institution already offers RDM services (via SURF) and if this is the case, we **refer the PI towards the institution**. This process is not yet automated. The challenge here is that the institutions mostly do not have a contact person or address yet.

c. Research Drive has a **dashboard** where the institute can manage users, project folders and internal contracts for certain amounts of storage. Within the dashboard there are also users with several roles associated with certain capabilities. Soon the dashboard will have reporting

ADVICE

The most regular contact with researchers takes place as part of the use of computing facilities. It is therefore logical to inform researchers there about data management. This is also the phase where the research projects are processing the data. Via the national HPC infrastructure, direct contact with researchers exists. At the same time the HPC computing centres are in contact with universities and research institutes. This differs per country, but still there are possibilities to make pragmatic connections for data management.

capabilities where the management roles in the dashboard have overviews of activities that users perform that fits their oversight role. This functionality was added after collecting feedback from the institutes using the dashboard. For the implementation of this new functionality, we have collected legal advice to ensure it is **GDPR-compliant**.

3 OUTPUTS / OUTCOMES

Institutions have been positive about collaboration and co-linking services. Many want an overview of the usage of data storage and management solutions. Some don't want extra costs (and certainly not when funded option is available).

Universities of Applied Sciences are the most active in organizing data management solutions for their researchers, actively transferring them all to for example Research Drive to get all the data in one place. But the researchers of Universities of Applied Sciences do not make a lot of use of computing yet, so there is less opportunity to actually push them from the grant towards their institution.

For SURF the challenges in connecting compute to data have mostly concerned processes, teams, systems and mandates. When connecting researchers to their central institution the main issues are process challenges i.e. having up-to-date details of what is available and who can help.

Further steps

- Assigning PIDs to steps in the compute process automatically, instead of by hand and just at the end.
- Implementing SRAM for every compute user, (authentication via institution).
- Integration of an iRODS (virtual RDM layer) platform community edition, to be able to 'move' data between different compute systems, different compute grant periods and nationally funded and institutionally funded services.

FOR MORE INFORMATION
Portfolio data services: https://edu.nl/bh8j6
Surf Research and IT: https://www.surf.nl/en/research-it

Authors

Mari Tinnemans and Sylvia Kuijpers

Credits and acknowledgement

NWO Allocation Computing Time department, Ron Trompert, Claudia Behnke, Peter Hinrich, and more.

GÉANT

GEANT.ORG @GEANTNEWS @GEANTCOMMUNITY /COMPANY/GEANTNETWORKS

SURF

SURF.NL

As part of the GÉANT 2020 Framework Partnership Agreement (FPA), the project receives funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 856726 (GN4-3).