‘Towards globally connected research data cloud infrastructure’

Jakub Mościcki[2], Gavin Kennedy[1]

Guido Aben[1]

[2] CERN, Geneva, Switzerland, kuba@cern.ch
[1] AARNet, Australia, firstname.lastname@aarnet.edu.au

INTRODUCTION

Over the past ~4 years, research file storage services based on desktop synchronization and sharing have taken off in earnest – in Australia, AARNet’s CloudStor is an example of this trend, but the same pattern is visible overseas, where similar services have become successful on a national level, gradually becoming an indispensable element of daily workflows for hundreds of thousands of users including researchers, students, scientists and engineers. The industry term for these services is “enterprise file synch&share”, or EFSS, which we’ll use throughout the remainder of this text.
Country by country, these services are typically operated and funded by major e-infrastructure providers, NRENs (National Research & Education Networks) and major research institutions; a leading example of a research establishment operating an EFSS service at web scale is CERN, the European Organisation for Nuclear Research, with their CERNBox platform. These services have seen very eager uptake by the market and have become an immediate bottom-up success. Likewise, the international community deploying and developing these services grew bottom up, and found its home at the yearly CS3 conferences (cs3community.org). Coordination is beginning to emerge from this forum, but entirely on a voluntary basis, without the presence of an umbrella body or a central mandate. At present though, these country-based services remain largely unconnected to each other, and some degree of wheel reinvention has been happening.

CATCHING UP WITH SUCCESS

Well-used data services are now dotted across the science landscape, but they are still run in isolation from each other. The obvious thing to do, which the systems and user bases are ready for, is to interlink these various systems and services and thus arrive at a consolidated view of both the data held within, as well as the user community seeking to collaborate through these systems. Such a joined-up system would enable much greater global collaboration and would more easily serve as a deployment platform for open science activities that benefit from single (virtual) deployment or a single world-view (e.g., citation, archival, metrics). It would save any user of system #1 of searching in vain for a collaborator or dataset present only on system #2. In short, this would improve the Findable, Accessible, Interoperable and Reusability (FAIRness) of data in active research projects, for all sites involved.

OOPS WE'RE FUNDED – IT'S DELIVERY TIME!

Fortunately, the European Commission agreed with the concept and, just recently, funded a consortium proposal (dubbed "CS3Mesh4EOSC") to build this interlinking system as part of the European Open Science Cloud (EOSC), with AARNet as an overseas partner. Important substrate services and protocols already exist and can be reused: the Open Cloud Mesh (OCM) protocol, co-developed by R&E networks under the GEANT banner, can already signal data shares between EFSS systems; and EduGAIN (deployed and supported in Australia by the AAF) can signal users and authorization levels between different countries' access federations.

This talk will introduce how CS3Mesh4EOSC will interconnect between its own nodes, and more importantly, how it will present the resultant service to the outside world, through a unified, stable API that's under development by the consortium. Examples of integration will be presented; federated compute at site #1 running on top of federated storage at site #2 (through Jupyter notebooks); taking data from EFFS store #1 and publishing it on the repository connected to EFFS store #2; universal, tiered search running on site #1 and accessing the indexed datasets on stores #2, #3 etc.

REFERENCES

[tbd]