Towards World-class Earth and Environmental Science Research in 2030: Will Today’s Practices in Data Repositories Get Us There?

Lesley Wyborn
National Computational Infrastructure, Canberra, Australia

Internationally Earth and environmental Science datasets have the potential to contribute significantly to resolving major societal challenges such as those outlined in the United Nations 2030 Sustainable Development Goals (SDGs). By 2030, we know that leading-edge computational infrastructures will be exascale (repositories, supercomputers, cloud, etc) and that these will facilitate realistic resolution of research challenges at scales and resolutions that cannot be undertaken today. Hence, by 2030, the capability for Earth and environmental science researchers to make valued contributions will depend on developing a global capacity to integrate data online from multiple distributed, heterogeneous repositories. Are we on the right path to achieve this?

Today, online, data repositories are a growing part of the research infrastructure ecosystem: their number and diversity has been slowly increasing over recent years to meet the demands that traditional institutional or other generic repositories can no longer satisfy. Although more specialised repositories are available (e.g., those for petascale volume data sets and domain specific long tail, complex data sets), funding for these specialised repositories is rarely long term.

Through initiatives such as the Commitment Statement from the Coalition for Publishing Data in the Earth and Space Sciences, publishers are now requiring that datasets that support a publication be curated and stored in a ‘trustworthy’ repository that can provide a DOI and a landing page for that dataset, and if possible, can also provide some domain quality assurance to ensure that data sets are not only Findable and Accessible, but also Interoperable and Reusable. But the demand for suitable domain expertise to provide the “I” and the “R” is far exceeding what is available. As a last resort, frustrated researchers are simply depositing the datasets that support their publications into generic repositories such as Figshare and Zenodo, which simply store the file of the data: rarely are domain-specific QA/QC procedures applied to the data.

These generic repositories do ensure that data is not sitting on inaccessible personal c-drives and USB drives, but the content is rarely interoperable. This can only be achieved by repositories that have the domain expertise to curate the data properly, and ensure that the data meets minimum community standards and specifications that will enable online aggregation into global reference sets. In addition, most researchers are only depositing the files that support a particular publication, and as these files can be highly processed and generalised they difficult to reuse outside of the context of the specific research publication.
To achieve the ambition of Earth and environmental science datasets being reusable and interoperable and make a major contribution to the SDGs by 2030, then today we need:

More effort and coordination in the development of international community standards to enable technical, semantic and legal interoperability of datasets;

To ensure that publicly funded research data are also available without further manipulation or conversion to facilitate their broader reuse in scientific research particularly as by 2030 as we will also have greater computational capacity to analyse data at scales and resolutions currently not achievable.