Scientific Papers

Robert Huber, Claudio D'Onofrio, Anusuriya Devaraju, Jens Klump, Henry W. Loescher, Stephan Kindermann, Siddeswara Guru, Mark Grant, Beryl Morris, Lesley Wyborn, Ben Evans, Doron Goldfarb, Melissa A. Genazzio, Xiaoli Ren, Barbara Magagna, Hannes Thiemann,

06 Feb 2021 .
Journal/Conference:

Highlights

- Uniform procedure to make environmental data analysis ready.
- Roadmap for related technology harmonization among research infrastructures.
- Substantial standardization effort required to enable interdisciplinary data re-use.
- Specialized software libraries can bridge standardization gaps.
- Use of common web standards can catalyse progress.

Abstract

When researchers analyze data, it typically requires significant effort in data preparation to make the data analysis ready. This often involves cleaning, pre-processing, harmonizing, or integrating data from one or multiple sources and placing them into a computational environment in a form suitable for analysis. Research infrastructures and their data repositories host data and make them available to researchers, but rarely offer a computational environment for data analysis. Published data are often persistently identified, but such identifiers resolve onto landing pages that must be (manually) navigated to identify how data are accessed. This navigation is typically challenging or impossible for machines.

This paper surveys existing approaches for improving environmental data access to facilitate more rapid data analyses in computational environments, and thus contribute to a more seamless integration of data and analysis. By analysing current state-of-the-art approaches and solutions being implemented by world-leading environmental research infrastructures, we highlight the existing practices to interface data repositories with computational environments and the challenges moving forward.

We found that while the level of standardization has improved during recent years, it still is challenging for machines to discover and access data based on persistent identifiers. This is problematic in regard to the emerging requirements for FAIR (Findable, Accessible, Interoperable, and Reusable) data, in general, and problematic for seamless integration of data and analysis, in particular. There are a number of promising approaches that would improve the state-of-the-art. A key approach presented here involves software libraries that streamline reading data and metadata into computational environments. We describe this approach in detail for two research infrastructures. We argue that the development and maintenance of specialized libraries for each RI and a range of programming languages used in data analysis does not scale well.

Based on this observation, we propose a set of established standards and web practices that, if implemented by environmental research infrastructures, will enable the development of RI and programming language independent software libraries with much reduced effort required for library implementation and maintenance as well as considerably lower learning requirements on users. To catalyse such advancement, we propose a roadmap and key action points for technology harmonization among RIs that we argue will build the foundation for efficient and effective integration of data and analysis.

The paper is structured as follows: Section 2 (Survey) presents the conducted survey, with a short description of the surveyed RIs and a description of the activities conducted to understand if and how the RIs support machine discovery of data access, given an identifier (e.g., digital object identifier (DOI)). Building on the survey, sections 3 (Solutions) and 4 (Discussion) present and discuss state-of-the-art solutions. In Section 5 (Roadmap), we suggest that the solutions can inspire a concerted technology harmonization among RIs that would enable the development and maintenance of RI and programming language independent solutions for data-analysis integration. Section 6 (Conclusions) closes this work with final remarks.

Read the full paper here

1,557 Read
<?php// print render($content['links']); ?>