ARTICLE IN PRESS

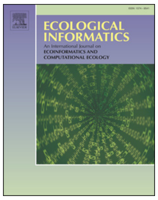
ECOINF-00701; No of Pages 10

Ecological Informatics xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Development of a cloud-based platform for reproducible science: A case study of an IUCN Red List of Ecosystems Assessment

Siddeswara Guru [a,*], Ivan C. Hanigan [a], Hoang Anh Nguyen [b], Emma Burns [c,d], John Stein [c], Wade Blanchard [c], David Lindenmayer [c,d], Tim Clancy [a]

[a] *Terrestrial Ecosystem Research Network, University of Queensland, St Lucia, 4072, Australia*
[b] *Research Computing Centre, University of Queensland, St Lucia, 4072, Australia*
[c] *Fenner School of Environment and Society, The Australian National University, Canberra, ACT 2601, Australia*
[d] *Long Term Ecological Research Network, Terrestrial Ecosystem Research Network, Australia*

ARTICLE INFO

Article history:
Received 24 March 2016
Received in revised form 15 August 2016
Accepted 20 August 2016
Available online xxxx

Keywords:
IUCN ecosystems assessment
Kepler workflow
Cloud computing
Platform as a service

ABSTRACT

One of the challenges of computational-centric research is to make the research undertaken reproducible in a form that others can repeat and re-use with minimal effort. In addition to the data and tools necessary to re-run analyses, execution environments play crucial roles because of the dependencies of the operating system and software version used. However, some of the challenges of reproducible science can be addressed using appropriate computational tools and cloud computing to provide an execution environment.

Here, we demonstrate the use of a Kepler scientific workflow for reproducible science that is sharable, reusable, and re-executable. These workflows reduce barriers to sharing and will save researchers time when undertaking similar research in the future.

To provide infrastructure that enables reproducible science, we have developed cloud-based Collaborative Environment for Ecosystem Science Research and Analysis (CoESRA) infrastructure to build, execute and share sophisticated computation-centric research. The CoESRA provides users with a storage and computational platform that is accessible from a web-browser in the form of a virtual desktop. Any registered user can access the virtual desktop to build, execute and share the Kepler workflows. This approach will enable computational scientists to share complete workflows in a pre-configured environment so that others can reproduce the computational research with minimal effort.

As a case study, we developed and shared a complete IUCN Red List of Ecosystems Assessment workflow that reproduces the assessments undertaken by Burns et al. (2015) on Mountain Ash forests in the Central Highlands of Victoria, Australia. This workflow provides an opportunity for other researchers and stakeholders to run this assessment with minimal supervision. The workflow also enables researchers to re-evaluate the assessment when additional data becomes available. The assessment can be run in a CoESRA virtual desktop by opening a workflow in a Kepler user interface and pressing a “start” button. The workflow is pre-configured with all the open access datasets and writes results to a pre-configured folder.

1. Introduction

Computation and software tools are used extensively in scientific analyses and the ensuing scientific publications are intended to disseminate data analyses and interpret their results. However, journal publications largely fail to provide sufficient information to generate verifiable knowledge (Donoho et al., 2009). According to Pritsker (2012), it is estimated that only 10%–30% of published studies are repeatable, although Freedman et al. (2015) claim that 51%–89% of all published research is not reproducible, and that one in four data analyses contribute to this irreproducibility. Specifically, data analysis steps are often not explained adequately in the literature for an independent researcher to reproduce the results. To quantify the cost of reproducibility, in one of the experiments, Garijo et al. (2013) estimated that it took 280 h to reproduce the results of a computational biology paper and concluded that only expert researchers in the domain could do it (Garijo et al., 2013).

Poor reproducibility has generated considerable concern and debate within the scientific community (Yamada and Hall, 2015) (Errington et al., 2014), and among funding agencies (Budget, 2014) (Collins and Lawrence, 2014) and scientific journal publishers (Editorial, 2012) (Jasny et al., 2011). The prevailing concern is that research might lacks credibility and cannot be trusted, and that poor reproducibility of

* Corresponding author.
E-mail address: s.guru@uq.edu.au (S. Guru).

http://dx.doi.org/10.1016/j.ecoinf.2016.08.003
1574-9541/Crown Copyright © 2016 Published by Elsevier B.V. All rights reserved.

scientific research precludes building on previous studies and leads to the loss of scientific efficiency (Marcus, 2014).

The seriousness of this problem has caused US National Institutes of Health (NIH) to initiate a plan to enhance reproducibility (Collins and Lawrence, 2014), and it was estimated that US $28 billion per year is spent on preclinical research that is not reproducible (Freedman et al., 2015). Measures have also been taken to improve reproducibility in other disciplines such as life sciences (Collins and Lawrence, 2014) and psychology (Asendorpf et al., 2013). More generally, there have been calls for all scientific claims to be completely repeatable by independent investigators who 'address a hypothesis and build evidence for or against it' (Peng, 2011). However, it is important to note that the exact results need not be computed in a repeatable/replicated study. This is because replication involves probability, and if performed again with a different sample and a new set of measurement errors, some variance between experiments is expected.

To date, concepts such as reproducibility and replicability have been defined in a number of ways. According to Cassey and Blackburn (2006), reproducibility is where the information presented in the study can be used by a third party to obtain the same results. This definition distinguishes reproducibility from replicability, which indicates that 'a third party must be able to perform a study using identical methodological protocols and analyse the resulting data in an identical manner for independent verification' (Cassey and Blackburn, 2006). Leek and Peng (2015) define reproducibility as 'the ability to recompute the results' and replicability as 'the chance that another experimenter can independently perform the experiment and achieve consistent results' (Leek and Peng, 2015). Peng (2011) described reproducibility and replication in a reproducibility spectrum where full replication is the highest form of reproducibility, and suggest that sharing code and data will only lead to partial reproducibility (Peng, 2011). From an experimental science perspective, reproducibility is more about using the information provided by original researchers to generate results, and replicability is where researchers independently recreate the experiment and get similar results.

In essence, reproducibility provides assurance to readers that the claims made in a publication are valid (Mesirov, 2010), and therefore improves the credibility of science. However, the benefits go further because reproducibility will increase scientific advancement through greater transparency, which will enable others to understand the analysis steps and the assumptions and constraints of a given study in more detail. This will limit waste of research funds and human resources because researchers will be better able to build on previous work.

To improve reproducibility in science, several journals, including the Public Library of Science (https://www.plos.org) and Vadose Zone Journal (http://vzj.geoscienceworld.org/), now require authors of scientific papers to deposit their data in public repositories (Skaggs et al., 2015). This is the first step towards enabling reproducibility, but may not be enough to repeat the published research. The Journal of Biostatistics has a more comprehensive policy to promote reproducibility, requiring that any article published with data or a source code is given a "D" or "C" kite-mark, respectively, and an article that meets these reproducibility requirements will be given an "R" (Peng, 2011). One of the major challenges with this approach is that the onus is on the journal to ensure that papers are reproducible using the code and data provided by the authors. If analysis steps are complex, repeating the simulation could be a difficult and time-consuming task for journal editors and reviewers. Hence, although we agree with Peng (2011) that reproducibility should become the minimum standard for assessing the value of scientific claims, to achieve this goal consistently, the onus should be on the proponent of the research to make the data analysis and synthesis steps reproducible. Accordingly, researchers will need to be provided with appropriate resources, incentives, and assistance to make their data and code accessible for reproduction. In certain instances, it may not be possible to reproduce the results even with the required source code and data because computational environments to execute the source code may not be available and/or easy to recreate. Hence, a major barrier to reproducible research is the 'lack of an integrated infrastructure for distributing reproducible research to others' (Peng, 2011).

In this paper, we have taken a holistic approach to reproducibility, where a complete computational analysis processes is available in an executable platform, irrespective of the complexity of the scientific research. The executable platform is accessible to independent researchers in a computation environment so that analyses can be repeated with minimal effort by other researchers. This makes reproducibility a non-tedious task for other researchers and where appropriate, could assist the original researchers in the longer-term to repeat the computational processes at different time-intervals. We therefore consider computational research as reproducible when computational experiment $\boldsymbol{e}$ that was developed at time $\boldsymbol{t}$ on a hardware infrastructure $\boldsymbol{h}$ using data $\boldsymbol{d}$ is repeated at time $\boldsymbol{t^1}$ on hardware $\boldsymbol{h^1}$ using data $\boldsymbol{d^1}$ that is similar to $\boldsymbol{d}$ (Freire et al., 2012). Thus, we built infrastructure in which computational research $\boldsymbol{r}$ that was developed at time $\boldsymbol{t}$ on a hardware and software infrastructure $\boldsymbol{h}$ using data $\boldsymbol{d}$ was reproduced at time $\boldsymbol{t^1}$ on the same hardware and software infrastructure $\boldsymbol{h}$ using the same data $\boldsymbol{d}$. This infrastructure is accessible in a very easily consumable manner so that users don't require any special tools to access them. In addition, the analysis tools and datasets used in the original computation are conserved and accessible.

The present manuscript describes the development of the integrated cloud-based infrastructure that allows building and distribution of reproducible computational-centric research. To demonstrate the value of the approach, we repeated an International Union for Conservation of Nature (IUCN) Red List of Ecosystems Assessments study conducted by Burns et al. (2015), and reproduced published results and made the computational processes of the assessment accessible to other researchers. The IUCN Red List of Ecosystems (http://www.iucnredlistofecosystems.org) provides a framework to assess the vulnerability of ecosystems and is akin to the IUCN Red List of Threatened Species (http://www.iucnredlist.org/).

Our paper is structured as follows: Section 2 further identifies the requirements and challenges of reproducible research, and the benefits of using a scientific workflow and cloud computing infrastructure; Section 3 describes the design, development and operation of the present Collaborative Environment for Ecosystem Science Analysis and Synthesis; Section 4 provides a brief overview of the IUCN Red List of Ecosystems Assessment conducted by Burns et al. (2015), and the implementation of this assessment as a Kepler scientific workflow; Section 5 concludes the paper and is followed by a statement of necessary future work in Section 6.

2. Building blocks of reproducible computational research

With the advancement of computational science, almost all analyses are performed using software tools. Therefore, the possibility of making research reproducible is quite high. However, most complex analyses involve (1) multiple datasets, (2) various algorithms and analytical tools and (3) high performance computing and/or (4) distributed computing infrastructure. There are therefore complexities and challenges to verifying scientific claims, and to providing the computational infrastructure to enable this. To address these challenges, researchers have recommended best practices for performing computational science (Sandve et al., 2013; Stodden and Miguez, 2014) (Wilson et al., 2014) (Nosek et al., 2015) (Kenall et al., 2015), and although we endorse these recommendations, we believe that the importance of the execution environment for successfully repeating any published analysis is not given the prominence it deserves. For published research to be reproducible, the following building blocks should be available:

- Data: all datasets used in the analysis must be available in an accessible digital format.
- Analysis procedures and methods: a detailed description of all the