Chapter 5
Documenting Provenance for Reproducible Marine Ecosystem Assessment in Open Science

Xiaogang Ma
Rensselaer Polytechnic Institute, & University of Idaho, USA

Peter Fox
Rensselaer Polytechnic Institute, USA

Stace E. Beaulieu
Woods Hole Oceanographic Institution, USA

Massimo Di Stefano
Rensselaer Polytechnic Institute, USA & University of New Hampshire, USA

Linyun Fu
Rensselaer Polytechnic Institute, USA

Patrick West
Rensselaer Polytechnic Institute, USA

ABSTRACT

Open Science not only means the openness of various resources involved in a scientific study but also the connections among those resources that demonstrate the origin, or provenance, of a scientific finding or derived dataset. In this chapter, the authors used the PROV Ontology, a community standard for representing and exchanging machine-readable provenance information in the Semantic Web, and extended it for capturing provenance in the IPython Notebook, a software platform that enables transparent workflows. The developed work was used in conjunction with scientists' workflows in the Ecosystem Assessment Program of the U.S. NOAA Northeast Fisheries Science Center. This work provides a pathway towards formal, well-annotated provenance in an electronic notebook. Not only will the use of such technologies and standards facilitate the verifiability and reproducibility of ecosystem assessments, their use will also provide solid support for Open Science at the interface of science and ecosystem management for sustainable marine ecosystems.

INTRODUCTION

Open Science has been receiving significant attention in recent years (Nosek et al., 2015). The Open Science movement includes open access to publications (Harnad & Brody, 2004), open data (Uhlir & Schröder, 2007; Glaves, H., 2017, chapter 3 this book), open source software and web services (Hey &

DOI: 10.4018/978-1-5225-0700-0.ch005

Payne, 2015), open collection of physical samples used in research (Lehnert, Vinayagamoorthy, Djapic, & Klump, 2006), as well as Science 2.0 (Shneiderman, 2008) that uses social media to exchange ideas and facilitate collaborations among scientists. Reproducibility is one of the key points that Open Science addresses (Open Science Collaboration, 2013; Yaffe, 2015). Efficient organization of the data, software programs, and samples used in a publication allows the authors to replicate the research. Making those resources open will further allow readers of that publication to verify the reported findings and, possibly, to reuse the resources in new studies. Besides reproducibility of research, the Open Science movement is also important in "actionable science" and connecting science to policy and management decisions (Ma, Fox, Tilmes, Jacobs, & Waple, 2014a; Palmer, 2012; Reichman, Jones, & Schildhauer, 2011).

Making publications, data, code, and samples open does not mean sharing them as fragmented pieces. In the era of the World Wide Web, various resources shared online can be linked in innovative ways (Emile-Geay & Eshleman, 2013). The work of provenance provides both methodological and technological approaches to associate those resources of Open Science on the Web and make the lineage and workflow of research traceable. The literal meaning of provenance is "the origin of something". In scientific works, documenting provenance on the Web involves linking and presenting the 'graph' structure of a network of research activities, people, and organizations involved in the production of scientific findings. This network includes the supporting observations, datasets, models, and methods used to generate those findings. On the Web, provenance documentation can be achieved through four primary steps: categorization of resources, assigning a unique identifier to each resource, annotation of those resources, and linking among them (Ma et al., 2014a).

In this chapter, the authors present work on provenance documentation and transparent workflows in assessment reports for Large Marine Ecosystems (LMEs) (See Appendix A for a list of acronyms). The LME concept was developed by the U.S. National Oceanic and Atmospheric Administration (NOAA), in consultation with international partners, for conservation and management of living marine resources (Sherman, 1991). LMEs are large areas of ocean about 200,000 km^2 or greater, adjacent to the continents where primary productivity is generally higher than in open ocean areas (NOAA, 2011). To understand and compare the relative impacts of fishing efforts, habitat degradation, climate change, and natural cycles on marine ecosystems, many different types of data must be collected, integrated, and interpreted. Data collected for monitoring ecosystem indicators in an Integrated Ecosystem Assessment (Levin, Fogarty, Murawski, & Fluharty, 2009) may be as diverse as satellite-derived sea surface temperature, counts of zooplankton species from net tows, and landings data from commercial fisheries. Provenance documentation is increasingly recognized as important for data contributed to community repositories and archives, inclusive of NOAA data collections and products derived from these data (NOAA 2011).

As part of the efforts to employ cyberinfrastructure data technologies to facilitate quantitative analysis and synthesis in LME assessments, the work of provenance documentation presented in this chapter benefited significantly from Semantic Web technologies (Leadbetter, Cheatham, Sepherd and Thomas, 2017, chapter 4 this book). A key step in this work is to have a machine readable conceptual model (i.e., an ontology) for objects and relationships in the provenance for datasets and other products in the assessment of LMEs. A motivation for enabling machine-readable provenance for NOAA's assessment of LMEs is the U.S. Executive Order "Making Open and Machine Readable the New Default for Government Information" (Obama, 2013). In the remainder of this chapter, Section 2 provides more details about provenance representation and documentation in the Semantic Web, as well as an interactive environment - the IPython Notebook - for data analysis for LME assessment. Section 3 describes the ontology developed to capture provenance in the IPython Notebook. Section 4 presents how this