Chapter 1
Oceanographic Data Management:
Quills and Free Text to the Digital Age and "Big Data"

Justin J. H. Buck
British Oceanographic Data Centre, National Oceanography Centre, UK

Roy K. Lowry
British Oceanographic Data Centre, National Oceanography Centre, UK

ABSTRACT

In this chapter the authors look at technological issues within and the influence of technological developments outside oceanography on two very different, but interrelated, facets of ocean data: the labelling of parameters and the autonomous collection of oceanographic data. In particular, consideration is given to the influence of standards developed within computer science, such as Sensor Web Enablement (SWE), Resource Description Framework (RDF) and the Simple Knowledge Organisation System (SKOS). It is shown that the development of autonomy in oceanographic data collection is highly dependent on semantically sound data and metadata with presentation of related issues such as unambiguous data citation, resolution of systematic biases and integration of data from complementary but independently governed oceanographic observation programmes.

INTRODUCTION

Oceanographic data collection started some 500 years ago with scribes recording readings from tide poles using quill pens and parchment and data management was the curation of these parchments. During the first half of the last Century oceanographic data management comprised the curation of printed books. Very little had changed over hundreds of years.

However, from 1950 onwards the rate of change accelerated phenomenally with the birth of the digital age. Oceanographic data management became formalised as an international activity during the 1960s. Since then the technological environment has changed beyond all recognition, which has inevitably revolutionised data management practices in line with new usage of data.

DOI: 10.4018/978-1-5225-0700-0.ch001

The chapter will examine the history of a selection of data management threads, looking at the impacts on them of technological change and the consequences of this change in terms of data accessibility, data usability and the quality of science. There are many potential candidates for such a study. Some are obvious and directly linked to hardware developments, such as data volume constraints and data availability timescales. Others are more subtle involving the technological changes driving cultural changes, such as the development and adoption of standards (including controlled vocabularies) and the integration of data into the scientific literature. The ongoing increase of internet bandwidth has meant exposure and access of data via the web has become the norm in less than two decades.

The next stage in the journey will be to examine the current state of development of these threads, with particular reference to what data consumers are now able to do plus what they cannot. The needs of both operational and scientific users of data will be considered. Operational users readily accept crude data quality levels to achieve rapid delivery. This is fundamentally different to the high quality data needed for precise estimates of quantities to monitor climate change such as the heat content of the ocean where small errors in source data propagate into significant errors in scientific findings (Barker, Dunn, & Domingues, 2011; Pedro & Goni 2011).

Next the consequences of technological evolution will be addressed. When there are significant technology jumps in terms of storage or data delivery there is a risk that previously processed data will be left behind making it unavailable or inconsistent in terms of format or semantic description. Aggregate datasets spanning centuries or longer enable long term scientific analyses e.g. Roemmich, Gould, & Gilson (2012). Thus, avoiding the creation of legacy data issues should be part of the adoption of new technologies.

Finally, the Chapter will look at how these threads could develop further in the future using technologies such as the Semantic Web following Leadbetter, Cheatham, Shepherd and Thomas (2017, chapter 4 this book) and the solutions currently under development for the facets of the 'Big Data' problem. The 'Big Data' age will place stringent demands on data semantics, but if these demands are met it will make data accessible to new forms of science. However, if they are not then ambiguous or non-machine readable data descriptions may hinder progress. Further, there is the very real risk that unless cultural issues involving the move towards 'open data' are addressed then the technological progress and the resulting scientific progress might be stopped in its tracks.

This chapter will address these topics by examining two complementary case studies, the first being the development of oceanographic data semantics, before moving on to consider the rapid and ongoing revolution that is ocean observation by autonomous platforms. The latter of which identifies additional issues and topics such as data standards, lessons learnt and data citation developments.

BACKGROUND

Oceanographic and Operational Context

Data consumers for oceanographic data are broadening in scope beyond the research project based synoptic studies that tend to fund research cruises. A major additional consumer is the increasingly data hungry family of ocean forecast models that require low-grade versions of data less than 24 hours after observation. Up until the 1990s ocean data were only available once a research cruise had reached port and the physical media (tape, disk etc.) had been submitted to data centres. These data were then