Citation of oceanographic data sets by scientific publications: an exploratory analysis

Christopher W Belter (Chris.Belter@noaa.gov), LAC Group, NOAA Central Library

INTRODUCTION

Evaluation of scientific research is becoming increasingly reliant on publication-based bibliometric indicators. This trend could lead to the potential devaluation of other scientific activities, such as data curation, that do not necessarily involve formal publication.

Partially as a reaction to this trend, and partially in response to other factors, there is also a growing movement focused on the development of a formal mechanism for citing data sets in scientific publications. This movement aims to increase the transparency of scientific research, facilitate the validation of reported results, and credit the individuals and institutions involved in creating and curating scientific data sets.

This analysis aims to combine these trends by attempting to generate citation counts for three oceanographic data sets archived by NOAA's National Oceanographic Data Center (NODC). By counting citations to these data sets, I attempt to credit NODC archivists in bibliometric terms. In doing so, I hope to identify how scientific articles currently cite these data sets to better inform the discussion about developing a standard method of data citation.

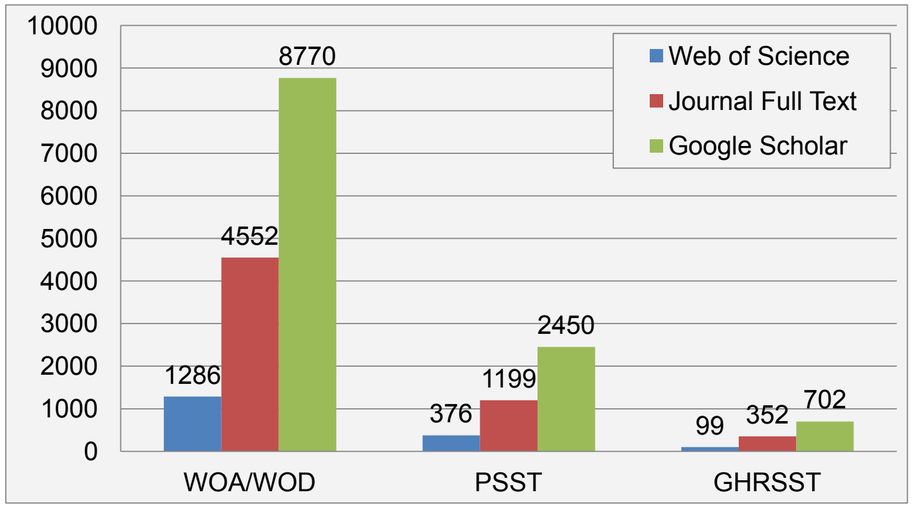
METHODS

In consultation with NODC, I select three highly-used data sets for analysis—the World Ocean Atlas and World Ocean Database (WOA/WOD), Pathfinder Sea Surface Temperature (PSST), and the Group for High Resolution Sea Surface Temperature (GHRSST)—and attempt to generate citation counts for each using three data sources: Web of Science (WoS), a selection of publishers' online full text repositories, and Google Scholar. In WoS, I search for mentions of these data sets in the title, abstract, keywords, and acknowledgements fields and also for citations to publications describing these data sets, where applicable. In the publishers' repositories and Google Scholar, I search for mentions of these data sets in the full text of scientific articles and compile the number of results retrieved.

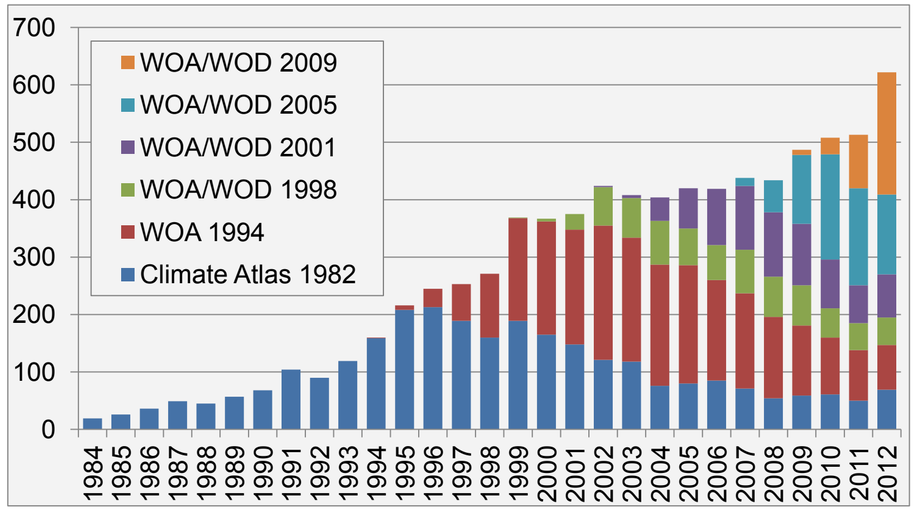
I then attempt a more comprehensive cited reference search in WoS to find citations to WOA/WOD. Since WOA/WOD was originally distributed as a formal publication, I hoped that formal citations to the data set would be more numerous. After initial cited reference searches for the title of the data set yielded unsatisfactory results, I performed cited reference searches for the authors and year of publication for each version of the data set and compiled the results to generate citation counts for each version of the set and for the set as a whole.

RESULTS

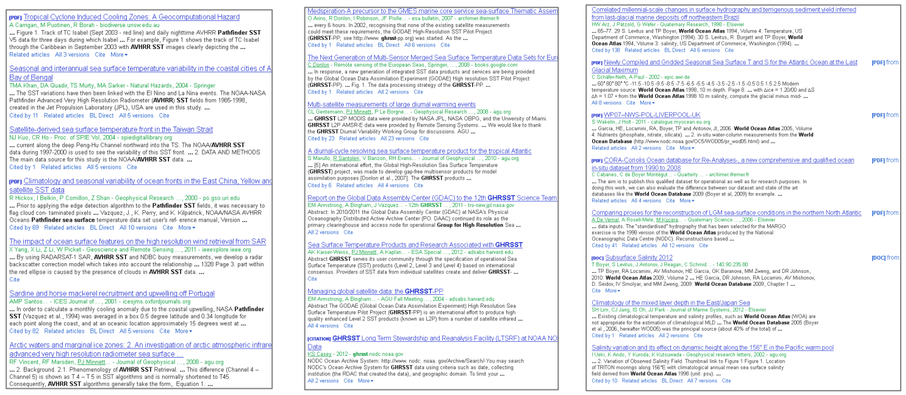
Citation Counts Derived from Three Data Sources



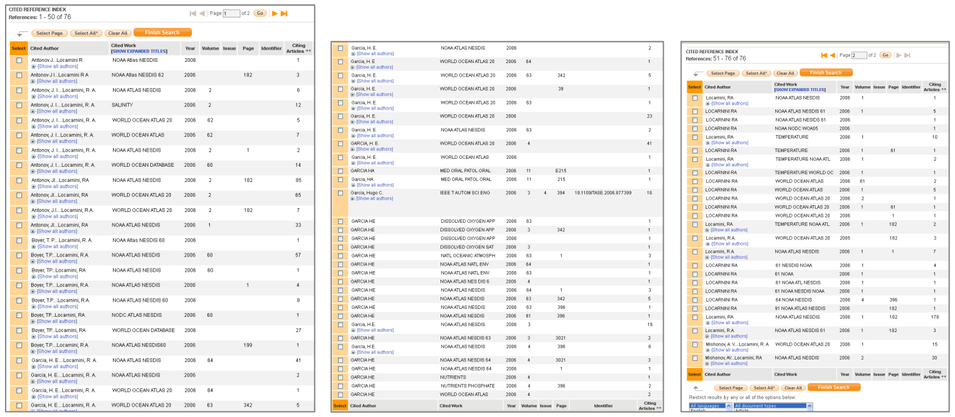
WOA/WOD Citation Counts Derived from Web of Science



Reference Variants from Google Scholar Results



Cited Reference Variants from WoS for WOA/WOD 2005



DISCUSSION

My results suggest that all of these data sets are highly cited. The Web of Science citation counts generated for WOA/WOD and PSST would, if they were counted as articles, rank these data sets in the top 1% for citation counts of all articles published in Oceanography in WoS in the same years. The citation counts generated through full text searching of publishers' sites, which was deliberately restrictive to improve search precision, would rank all three sets in the top 1%. My more comprehensive search results would rank the 1982, 1994, 2001, 2005, and 2009 editions of WOA/WOD as the most highly cited articles published in Oceanography in these years.

These results also suggest that the majority of references to these data sets occur in the full text of scientific articles, and not in the cited reference lists or acknowledgements sections of these articles. The citation counts generated through full text searching of publishers' sites are nearly four times higher than those generated in WoS, while the Google Scholar citation counts are nearly eight times higher than the WoS counts. These results call into question the appropriateness of using citation databases like WoS or Scopus to generate citation counts for these data sets.

Finally, my results suggest that these data sets are referenced in a wide variety of ways, even when a formal citation is suggested. I find over 200 variant methods of citing the 2005 edition of the WOA/WOD, despite the fact that a standard citation method is provided. This suggests that although developing a data citation standard is necessary, it is not sufficient to guarantee consistent citation to these data sets. Once a citation standard is developed, significant additional effort would seem to be necessary to inform scientists about the standard and to obtain commitments from journal reviewers and editors to enforce that standard in order to ensure that these data sets are properly and consistently cited.

Presented at the 2013 RDAP Summit, 4-5 April, Baltimore MD