Analyzing Data Citations to Assess the Scientific and Societal Value of Scientific Data

Robert R. Downs[2], Robert S. Chen[1] , Joe Schumacher[3]

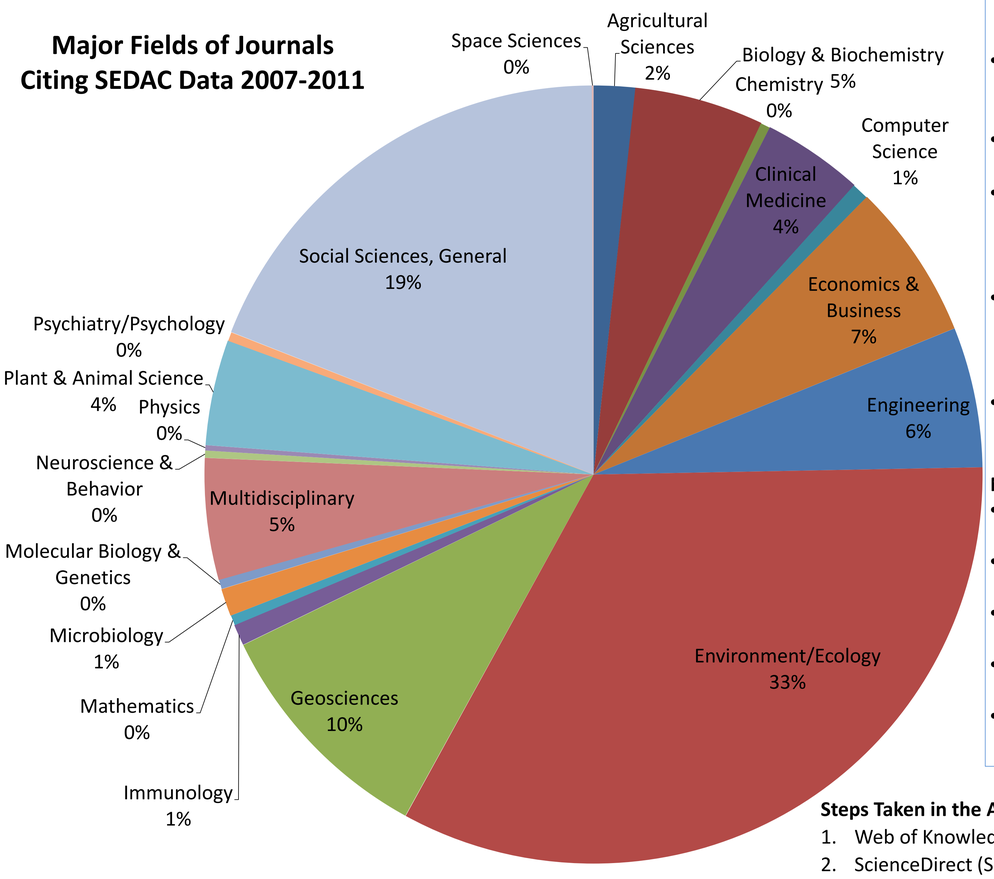
NASA Socioeconomic Data and Applications Center (SEDAC)
Center for International Earth Science Information Network (CIESIN)
Columbia University

[1] bchen@ciesin.columbia.edu
[2] rdowns@ciesin.columbia.edu
[3] jschumac@ciesin.columbia.edu

2013 Columbia University Research Data Symposium
February 27, 2013, New York, NY

Stakeholders in the creation, distribution, support, funding, and use of scientific data can benefit by assessing the value that the data have for society and science. For decades, the scientific community has used citations of articles in the published scientific literature as one of the primary measures for evaluating the performance and productivity of scientists, departments, institutions, and scientific disciplines. Similarly, citations of scientific data in the published literature may be useful for tracking and comparing the value of the scientific data and the contributions of individuals, projects, programs, and organizations to the data's development and use. Citation analysis can contribute to planning for future data collection, development, distribution, and preservation efforts. The release of new data citation indexes and more widespread adoption of unique data identifiers and automated attribution mechanisms have the potential to improve significantly the capabilities for analyzing citations of scientific data. In addition, ongoing developments in the systems and capabilities for disseminating data, along with education and workforce training on the importance of data attribution and on techniques for data citation, can improve practices for citing scientific data. Such practices need to lead not only to better aggregate statistics about data citation, but also to improved characterization and understanding of the impact of data use with respect to the benefits for science and society. Analyses of citations in the scientific literature were conducted for data that were distributed by an interdisciplinary scientific data center during a five-year period (1997 – 2011), to identify the scientific fields represented by the journals and books in which the data were cited. Secondary citation analysis also was conducted for a sample of scientific publications that used the data extensively to identify the potential impact of the data on the scientific fields represented by those journals. Furthermore, an initial analysis was conducted of citations that appeared in non-peer-reviewed publications and the popular media to assess the potential policy and educational impacts of these data. The initial results of these analyses demonstrate the significant challenges that remain for consistent, quantitative assessment of the value of scientific data to both science and society .

Major Fields of Journals Citing SEDAC Data 2007-2011



Results and Interpretation

- Found 325 different peer-reviewed journals containing 674 articles during 2007-11
- Largest group of citations in ecological and biological fields esp. related to biodiversity
- About one quarter of citations in the social sciences (including economics)
- Only about 10% of citations strictly in the geosciences, though many of the environment/ecology citations may reflect similar work
- Multidisciplinary use of SEDAC data is not restricted to the limited number of journals categorized as multidisciplinary or interdisciplinary
- More limited use of SEDAC data in engineering, computer science, and medical fields

Needs for Further Research

- Refine discipline categories to match SEDAC user community definitions
- Assess search strategies for underrepresented disciplines
- Characterize type and "quality" of citations, incl. journal impact factors
- Examine trends and differences by dataset or data collection
- Begin assessing policy/grey literature

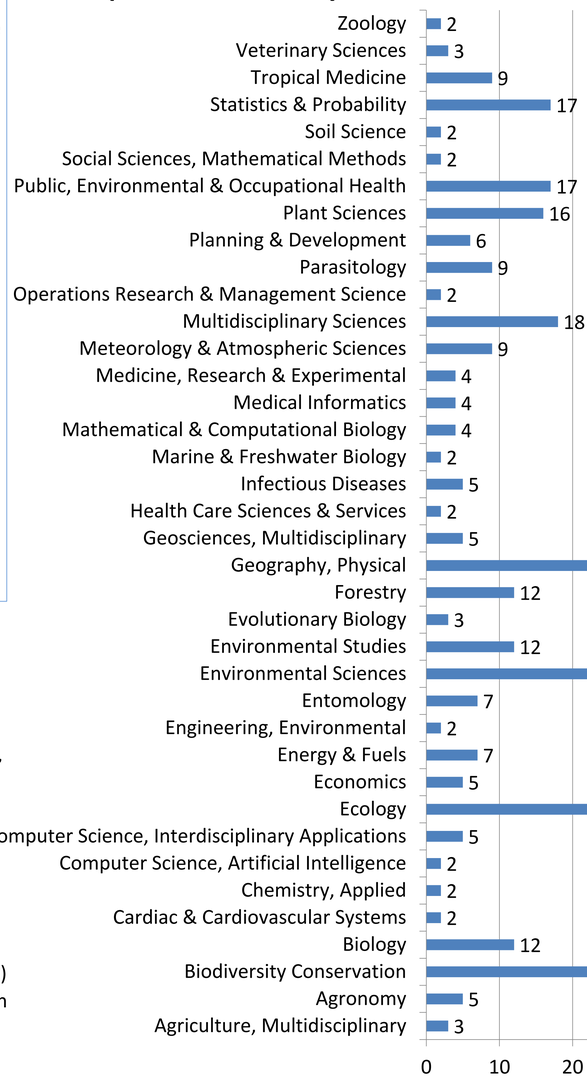
Identification of Works that Cite SEDAC Data

- Subscriptions to notification services (ScienceDirect, Springerlink, Scopus, etc.)
- Notifications from authors, publishers, or readers
- Searches of web and databases, for terms such as SEDAC, CIESIN, GPW, and "gridded population"

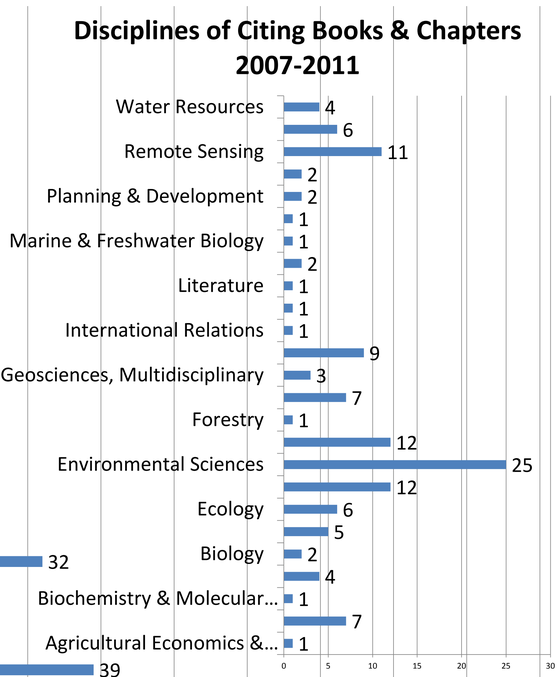
Steps Taken in the Analysis of Disciplines

1. Web of Knowledge (Web of Science Categories) for journals
2. ScienceDirect (Scopus subject categories) for journals
3. Publication titles (Subject terms)
4. Publication web sites (Subject categories, publication scope, summary)
5. Designation of disciplines
6. Journals (one or more disciplines for each journal)
7. Books (one discipline for each book)
8. Books with chapters (one discipline for each book)
9. Normalization of disciplines
10. Inconsistencies between discipline categories
11. Normalized disciplines using Web of Science Categories (5.3)
12. Major field identified from Thompson Reuters ScienceWatch Major Field of Science categories

Disciplines of Secondary Citations of Selected Articles Citing SEDAC Data



Disciplines of Citing Books & Chapters 2007-2011



The authors presented an earlier version of this poster at the 2012 AGU Fall Meeting in San Francisco, CA, on 6 December 2012.

Center for International Earth Science Information Network
Earth Institute | Columbia University

The authors appreciate the support received from the National Aeronautics and Space Administration (NASA) under Contract NNG08HZ11C for the Continued Operation of the Socioeconomic Data and Applications Center (SEDAC).