Disciplinary reach: investigating the impact of dataset reuse in the earth sciences

Tiffany C. Chao
Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
tchao@illinois.edu

ABSTRACT
In the realm of scholarly communication, scientific datasets are becoming more widely recognized for their scholarly and reuse value. However, given the investment toward maintaining and storing research data for long-term access, there is no clear strategy or metric for determining the reuse of research datasets. This study proposes a novel approach to track use and measure the impact of publically accessible datasets in scholarly publications through *disciplinary reach-* the number of unique journals and related subject categorizations in which articles are published. Using *affiliated publication(s)*, described by the author as the works identified by the dataset creator or curator related to a dataset, the principles underlying the bibliometric technique of citation analysis are leveraged and applied. Preliminary results show that for earth science datasets, affiliated publications were primarily cited in physical science and multidisciplinary journals, indicating these datasets may have an impact on a number of different research areas. Continued refinement of these approaches, measures, and the design will serve to broaden our understanding of the reuse potential of scientific data and their influence on advancing scholarship.

Keywords
Science data use, citations, data reuse, impact, journal

INTRODUCTION
Datasets are steadily emerging as commodities in scholarly communication, a position long held by journal publications (Davis & Vickery, 2007). Greater attention given to the value of scientific datasets is seen by some scholarly journal publishers, who now require that the datasets used in a study be submitted alongside the written work as a condition for publication (Whitlock, et al, 2010). The importance of datasets is also reflected in efforts to increase their public access through the formation of data repositories and cyberinfrastructures. Current web-based science data repositories span a number of scientific domains including astronomy, medicine, biology, and the geosciences, with many providing services for data submission, thereby promoting data sharing (Marcial & Hemminger, 2010). Within the United States, the National Science Foundation’s Office of Cyberinfrastructure (http://www.nsf.gov/dir/index.jsp?org=OCI) offers grant funding for the development of systems that facilitate the management, storage and long-term preservation of scientific data, particularly at a time where technological advancements in computing and instrumentation have enabled vast quantities of data to be generated for research and analysis.

The notion of sharing research data for public consumption is based on the premise that these data possess enduring value beyond their originally designed purpose (Pienta, Alter & Lyle, 2010). Through the continued use and reuse of research data, new questions and analytical techniques can be applied without the expense of collecting original data or duplicating previous efforts (Fienberg, Martin & Straf, 1985). As the dataset plays a direct role in formulating a researcher’s scholarly contribution, the subsequent impact of that work on the scientific discourse can be attributed back to the data. Within microarray research, Piwowar, Day & Fridsma (2007) identified a significant positive correlation between datasets that were openly available and the publication citation rate of the authors that contributed the dataset. By providing evidence connected to plausible incentives that resonate with scientific culture (i.e. increased publication citations for the researcher), there is the potential that an increase of sharing research datasets will lead to a rise in their reuse, which further necessitates a method to capture this usage within the scholarly literature. Based on this strategy, what makes a dataset amendable for reuse can be better understood in order to enable appropriate services to support this research practice.

Datasets and citation
One of the most well-known bibliometric techniques is citation analysis, the examination of references cited in the bibliographies of scholarly publications (De Bellis, 2010).
As a data source, citations are readily available, relatively static, and can be used by a number of analytical techniques such as citation counts, bibliometric coupling, or co-citation (Smith, 1981). They can also be utilized to evaluate new indicators and tools inline with bibliometric research. The approach and results of citation analysis have been steadily adopted and integrated into non-research venues including decision-making processes regarding promotion and tenure in academia as well as research policy development and distribution of grant funding (Weingart, 2005; Moed, 2005).

Citation analysis has the power to display the intellectual impact of journals, disciplines, and scientists, based on the premise that “influential works or scientists are cited more often than others. In this sense, citations reflect the relative impact and utility of a work, author, department, or journal’s publications within their larger scientific domains” (Meho & Rogers, 2008, p. 1712). Adapting this principle, an influential dataset should be cited more frequently than its counterparts. However, complications arise in practice when utilizing scholarly literature to identify dataset reuse. There are a number of motivating factors and intentions behind an author’s citing behavior which are not clearly defined in the use of a particular citation (Garfield, 1972). The inclusion of a certain citation is not always indicative of how the correlated publication was used; MacRoberts and MacRoberts (2010) found that a number of informational units in scientific publications from the field of biogeography considered to be influential contributions to the end scholarly product were oftentimes uncited, particularly data from secondary sources (i.e. databases) and other grey literature. Additional methods, such as content analysis, are needed to further elicit and explicate the meanings and relationship between the citation and its use (Bornmann & Daniel, 2008; Small, 1987).

For the quantitative dataset, a standard bibliographic citation is still under development and not commonly used in attribution for datasets used or reused within the published work by research communities (Altman & King, 2007). It is more common to see a journal publication or other documentation describing the dataset referenced, even though such a citation does not explicitly indicate what data are actually used (Parsons, Duerr & Minster, 2010). The dynamic nature of some sets of data, where information is periodically added and updated, presents the challenge of whether to consider these changes as versions of the original dataset and how the citation should then be represented (Green, 2009). Other attributes of a dataset, such as an assigned identifier (i.e. digital object identifier DOI), dataset title, submitter name, data collection article publication, and even queries, have been considered as channels to track dataset reuse, yet remain difficult to leverage due to their inconsistent use in the literature (Enriquez et al., 2010). While some processes of tracking reuse could potentially be automated, such as the retrieval of datasets based on DOIs, much of the work must still be manually completed.

Taking into consideration the obstacles inherent to tracing dataset use in the published literature, this study explores a potential measure of reuse by examining the relative impact of a dataset. More specifically, the following questions are posited to cultivate this measure: what bibliometric indicators can be employed or adapted to measure the scholarly impact of dataset reuse in journal publications? How can automated processes be leveraged to track dataset reuse in journal publications? A semi-automated approach is proposed that captures the citation history of affiliated publications, those works identified by the dataset creator or curator related to a dataset, in order to measure its relative impact. This study is also a reflection on the effectiveness of methodological approaches and techniques implemented to collect results for analysis.

**METHODS**

The methodological design was developed from a preliminary exploration of what type of information was needed and readily accessible to evaluate dataset use and the processes applied. Automated approaches were integrated when possible in the collection of publications and related information. Detailed in this section are the steps taken to identify datasets and retrieve affiliated publications, assess the disciplinary reach of the dataset, and evaluate the automated tools used.

For this study, the scientific domain of ‘earth science’ was selected due to known publically available datasets with corresponding publications. A sample of 6402 datasets was derived from the ‘Atmosphere’ dataset collection at NASA’s Global Change Master Directory (GCMD) (http://gcmd.nasa.gov/), a web-accessible public metadata inventory of earth science datasets. Each dataset metadata record had a field for ‘publications’, defined as “key bibliographic citations pertaining to the data set” (DIF, 2010). These affiliated publications to the dataset provided the basis for citation analysis. Similar to literature-based citation, the motivation and intent behind citing another publication is not always clear (Small, 1987) and in this case, the relationship between dataset and publication is also not clearly defined.

**Dataset identification and publication citation retrieval**

There are several key points from which data for analysis of impact need to be collected. Dataset ‘451’, an example of an average case from the sample, is used to illustrate what data were retrieved from the publication citations affiliated with the dataset (see Figure 1). The numbers represented in Figure 1 correspond to the following steps:

1) Data collection occurred from mid-March to April 2011. In order to retrieve dataset records from GCMD and extract affiliated publications, a customized Perl (v5.8.8) script was developed to harvest the webpage where the record was presented. For each dataset record, the listed publications were automatically extracted and assigned a unique
identifier. Publications retrieved not only included journal articles but also technical reports, theses, dissertations, conference proceedings, workshop reports, and links to external websites. Each publication citation was then parsed based on last name of the first author, title of publication, year of publication, and title of the journal in which the article was published. This parsed bibliographic citation is used to create a search query for Google Scholar.

In the example, the publically accessible dataset has the identifier ‘451’ along with four affiliated publications (A-D), all of which in this case, are from journals.

2) The citation search query for each affiliated publication was automatically inputted into Google Scholar (http://scholar.google.com/) to retrieve the number of citations received. Google Scholar was selected due to its breadth of literature coverage compared to Scopus or ISI Web of Science (Meho & Rogers, 2007; Kousha & Thelwall, 2008), a vital component given the varied nature of publications affiliated with the earth science datasets. The ‘cited by’ number given by Google Scholar provides an estimate for the citations accrued by an affiliated publication; this number if graphically shown by the total number of arrows extending from the ‘affiliated publication’ in Figure 1. These ‘cited by’ publications are designated as secondary publications, where any bibliographic information provided by Google Scholar (i.e. authors, title of publication, year of publication, venue of publication) was captured for each work.

3) In order to collect the necessary data for assessing impact, the number of citations received by each secondary publication was acquired. The ‘cited by’ URLs from the Google Scholar record were manually retrieved for each secondary publication. These URLs were then used with a second customized Perl script to acquire the static Google Scholar pages and parse the number of citations for each of these secondary publications. This final set of data collected only contains the number of citations received by secondary publications.

From the example in Figure 1, the one secondary publication from affiliated publication (A) was cited once, affiliated publication (B) was cited two times, affiliated publication (C) was cited six times, and affiliated publication (D) was cited 50 times.

The Figure 1 example portrays affiliated publication (A) was cited once, affiliated publication (B) was cited two times, affiliated publication (C) was cited six times, and affiliated publication (D) was cited 50 times.

From the example in Figure 1, the one secondary publication from affiliated publication (A) was cited three times and the two secondary publications from affiliated publication (B) were cited three and two times, respectively. Affiliated publication (C) had six secondary publications, cited four, two, two, two, zero, and zero times, respectively. Lastly, each of the 50 secondary publications from affiliated publication (D) was cited as many as 49 times or not at all. The groupings of arrows extending from ‘secondary publications’ represent the number of citations obtained (arrows) by each secondary publication (group); ‘0’ in place of an arrow indicates that the secondary publication has not been cited.

Figure 1. Overview of the data collection process through example dataset 451
Impact and H-index
As a broad measure of impact, the total number of citations received by affiliated publications was counted. Datasets with multiple affiliated publications present an inherent bias to this approach because there are more publications available to cite and a greater opportunity to accrue citations. The H-index proposes an alternative approach to measuring impact and productivity that considers the number of publications produced over a given period of time, the number of citations received by each publication, and accounts for highly cited articles as a bias in relation to less frequently cited papers from the same person (Hirsh, 2005: see paper for disadvantages of the index). Based on the premise that an H-index of 10 indicates 10 papers have at least 10 citations each, an initial calculation of this index was manually derived from the number of citations received by an affiliated publication. While the application of the H-index was originally determined at the author level, other scholarly avenues such as journals, have utilized this index (Braun, 2006).

A second contributing component for determining impact is the diversity of citations received by an affiliated publication outside of the immediate domain of research. This diversity or reach could be established by publication authors or the country of origin but for this study, disciplinary reach was decided by the journal of the publication and corresponding subject area based on categories established by Scopus (see: http://help.scopus.com/robo/projects/schelp/h_subject_categories.htm). Such subject classifications have been used in studies of interdisciplinary scientific research and are shown to evolve as new research areas emerge and disciplinary boundaries shift (Morillo, Bordons, & Gómez, 2003). Based on the number and distribution of unique subject areas identified for the affiliated and secondary publications, a relative impact measure for the datasets can be inferred.

Evaluation of automated processes
Throughout the process of data collection and analysis, general observations and patterns were recorded. The data output from two pairs of tools 1) the Perl scripts created to harvest the dataset records and parse the publications 2) the Perl scripts developed to harvest and parse Google Scholar publication citations, and the products produced were reviewed. These notes provide the basis to inform the refinement of customized scripts while also revealing hidden obstacles of the overall approach design, including the use of tools and measures selected for analysis.

PRELIMINARY RESULTS AND DISCUSSION
In the pursuit of understanding how the impact of dataset reuse can be measured through published works, this section presents preliminary results on the impact measures adapted and developed from the bibliometric literature (i.e. H-index and disciplinary reach) and the functionality of automated processes integrated in the data collection design to track dataset reuse.

Of the 6402 dataset records collected, a subsample of 600 datasets was identified for analysis based on the parameters of publication availability and date of public release. Nearly 60% of these data set records had two or more affiliated publications, with 27 as the largest count. Since the GCMD metadata records have update and modification capabilities, the addition of new affiliated publications is entirely possible from the time the data for this study was collected. New datasets are also made available on regular basis so changes in counts are expected in the future study. As depicted in Figure 1, the number of affiliated publications for a dataset differs from one to the next and the number of times each affiliated publication was cited also varies; the example portrayed in the Figure is also quite telling of how quickly and how much citation data was accumulated through this process.

Dataset impact and reach
The measure of disciplinary reach for a dataset was

<table>
<thead>
<tr>
<th>Top Ten Journals for Affiliated Publications</th>
<th>% of affiliated publications in journal</th>
<th>Top Ten Journals for Secondary Publications</th>
<th>% of secondary publications in journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal of Climate</td>
<td>10.75%</td>
<td>Journal of Climate</td>
<td>8.57%</td>
</tr>
<tr>
<td>Climate Dynamics</td>
<td>9.25%</td>
<td>Climate Dynamics</td>
<td>6.81%</td>
</tr>
<tr>
<td>Nature</td>
<td>6.00%</td>
<td>Geophysical Research Letters</td>
<td>6.48%</td>
</tr>
<tr>
<td>Science</td>
<td>3.92%</td>
<td>Climatic Change</td>
<td>4.31%</td>
</tr>
<tr>
<td>Tellus B</td>
<td>3.33%</td>
<td>International Journal of Climatology</td>
<td>4.26%</td>
</tr>
<tr>
<td>Bulletin of the American Meteorological Society</td>
<td>2.92%</td>
<td>Science</td>
<td>3.55%</td>
</tr>
<tr>
<td>Geophysical Research Letters</td>
<td>5.17%</td>
<td>Nature</td>
<td>2.96%</td>
</tr>
<tr>
<td>IEEE Trans. on Geoscience and Remote Sensing</td>
<td>2.58%</td>
<td>Quaternary Science Reviews</td>
<td>2.53%</td>
</tr>
<tr>
<td>Journal of Applied Meteorology</td>
<td>1.58%</td>
<td>Earth and Planetary Science Letters</td>
<td>2.38%</td>
</tr>
</tbody>
</table>

n= 1200                                      n= 30,000

Table 1. Comparison of ten most cited journals for affiliated and secondary publications
are considered to be ‘physical science’, the other areas journals were affiliated with a dataset and those subsequent publications produced that cited the dataset article. Among the top three journals with the most citations were consistent across both sets of publications. Among the ten most cited journals, there was also very little variation in the subject areas of journals where publications reside. While half of the journals that affiliated publications were published in are considered to be ‘physical science’, the other half of journals cross at least two subject areas, with two journals ‘Nature’ and ‘Science’ noted as ‘multidisciplinary’ in content. The number of multi-subject journals increases with secondary publications, for instance, the International Journal of Climatology bridges the areas of ‘life science’, ‘physical science’ and ‘social science’. This expansion suggests that the earth science datasets have a reach or impact across several scholarly domains.

Excluding the top ten journals for both the affiliated publications (n= 1200), nearly 60% of them were published in ten journals, even though 221 unique titles were identified (see Table 1). A similar pattern is observed for secondary publications (n= 30,000) and the journals in which they tend to be published. Approximately 6225 unique titles were identified but about 50% of the secondary publications are accounted for within ten journals (see Table 1). Journals and their assigned subject areas were compared between affiliated publications and their secondary publications using those standardized classifications established by Scopus. These subject areas are detailed in Table 2.

Analysis
In general, there were similarities between publications affiliated with a dataset and those subsequent publications produced that cited the affiliated dataset article. The top three journals with the most citations were consistent across both sets of publications. Among the ten most cited journals, there was also very little variation in the subject areas of journals where publications reside. While half of the journals that affiliated publications were published in are considered to be ‘physical science’, the other half of journals cross at least two subject areas, with two journals ‘Nature’ and ‘Science’ noted as ‘multidisciplinary’ in content. The number of multi-subject journals increases with secondary publications, for instance, the International Journal of Climatology bridges the areas of ‘life science’, ‘physical science’ and ‘social science’. This expansion suggests that the earth science datasets have a reach or impact across several scholarly domains.

Excluding the top ten journals for both the affiliated publications (n= 211) and secondary publications (n= 6215), the overall reach of the dataset goes beyond physical sciences to other scholarly domains such as the arts and humanities and social sciences. Secondary publications were cited in the Journal of Anthropological Archaeology (cited three times), classified as a ‘social science’ journal while other publications were seen in Journal of World History (cited three times), a ‘arts and humanities’ journal. The variety of journals for secondary publications may also account for the decrease of nearly 3% from affiliated to secondary publications cited in Nature; there were slight shifts in the order of the ten most highly cited journals for secondary publications but none as apparent as seen with Nature. Although more transparent adoption of datasets into other domain areas for scholarly use may not be so prevalent to warrant a top ten cited journal, it is this work at the fringes of disciplinary boundaries that results in innovative output with potentially significant changes for multiple scholarly areas (Cummings & Kiesler, 2005).

H-index results
The initial calculation of the H-index for affiliated publications citations is depicted in Figure 2. Approximately 41% of datasets had an H-index of ‘1’ suggesting that these earth science datasets may have a

<table>
<thead>
<tr>
<th>Top Ten Journals for Affiliated Publications</th>
<th>Scopus subject area</th>
<th>Top Ten Journals for Secondary Citations</th>
<th>Scopus subject area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal of Geophysical Research</td>
<td>Life/Physical Science</td>
<td>Journal of Geophysical Research</td>
<td>Life/Physical Science</td>
</tr>
<tr>
<td>Journal of Climate</td>
<td>Physical Sciences</td>
<td>Journal of Climate</td>
<td>Physical Sciences</td>
</tr>
<tr>
<td>Climate Dynamics</td>
<td>Physical Sciences</td>
<td>Climate Dynamics</td>
<td>Physical Sciences</td>
</tr>
<tr>
<td>Nature</td>
<td>Multidisciplinary</td>
<td>Geophysical Research Letters</td>
<td>Physical/Social Sciences</td>
</tr>
<tr>
<td>Science</td>
<td>Multidisciplinary</td>
<td>Climatic Change</td>
<td>Physical Sciences</td>
</tr>
<tr>
<td>Tellus B</td>
<td>Life/Physical Science</td>
<td>International Journal of Climatology</td>
<td>Life/Physical/Social Science</td>
</tr>
<tr>
<td>Bulletin of the American Meteorological Society</td>
<td>Physical Sciences</td>
<td>Science</td>
<td>Multidisciplinary</td>
</tr>
<tr>
<td>Geophysical Research Letters</td>
<td>Physical Sciences</td>
<td>Quaternary Science Reviews</td>
<td>Life/Physical Science</td>
</tr>
</tbody>
</table>

Table 2. Comparison of subject classification for ten most cited journals between affiliated and secondary publications
relatively low impact as scientific research contributions. The use of the H-index calculation did account for affiliated publications with extreme citation counts; 9% of datasets had affiliated publications with over 1,000 citations yet the highest H-index was 15.

Looking more closely at this “influential” dataset with an H-index of 15, the earliest affiliated publication dates to 1980 and the paper with the most citations (2324) dates to 1997. Publication dates were more varied for affiliated works with an H-index of one, ranging from 1986 to 2002. The six-year difference may be a contributing factor for a higher H-index, however, there is not enough evidence at this point to conclude that an earlier publication date is a definitive factor of datasets with high H-indices being more influential than their counterparts. There are a number of additional avenues for analysis based on the data collected for the H-index, including a more in-depth exploration of patterns that may indicate a dataset’s reuse potential.

Based on these initial findings, the measure of impact for dataset reuse can be communicated at both the individual dataset level as seen with the H-index and at the collective domain level demonstrated with disciplinary reach. Ideally, both measures of impact would be used for an individual dataset. Given the subject areas that were available, it is important to explore other subject area classifications outside of Scopus to further refine this definition of ‘reach’, especially given the diversity of ‘publications’ affiliated with a dataset. While the broad categories allowed for more straightforward classification of journals, these categories were also a deterrent since they overshadowed the subdisciplinary-level or specialized areas of study; for example, the ‘physical science’ category in Scopus ranges from chemical engineering to physics and astronomy which each also encompasses a number of subdisciplinary fields. Overall, the disciplinary reach of affiliated and secondary publications did reveal patterns and trends that could be used to indicate areas of study and research utilizing these datasets. Combined with the H-index calculation, which accounts for extreme citation counts, one dimension of how impact of dataset reuse can be assessed is illuminated.

**Process evaluation**

This methodological design employed several automated techniques to ascertain data for analysis. The output from customized scripts revealed inconsistencies in both the metadata record for the dataset in GCMD and Google Scholar publications. Published works parsed from the metadata record did not follow a consistent style of bibliographic citation despite standard fields designated for the metadata structure. For instance, identifying the year of publication required several common citation patterns to be included in the script but even with this precaution, approximately 16% of publication years were undetected or incorrectly identified (i.e. ‘7602’, ‘4697’). Part of this discrepancy can be accounted for by publications that were ‘accepted’ or ‘in press’ (1%) with no known date given, or by publications for which a date was not recognized (78%). Other sources of error were connected with publication titles containing a four or more numeric character string where the actual date of publication appeared subsequently in the citation (21%).

For the parsed publication citations from Google Scholar, barriers were witnessed in the completeness of records represented on the web pages. In the extraction of journal titles from publication citations, the level of consistency greatly varied for how the journal was listed. This was seen with the ‘Journal of Climate’: variations on this title
included ‘J. Climate’, ‘J. Clim.’, and ‘J. Clima…’; the full citation then needed to be retrieved, verified and normalized manually. This process was necessary to execute for all journal titles and in this case, no significant modifications could be made to the parser script. Particularly for non-traditional publications such as workshop reports or dissertations, bibliometric standards for citation varied even more.

An additional manual procedure undertaken examined the citation counts (‘cited by’) from Google Scholar and that they matched the number of affiliated publications reported. The year of publication for these affiliated works was also noted; it would not be plausible for an affiliated publication to be published and available prior to the article it was citing. These checks were conducted intermittently and should be more systematized in the future in order to maintain and improve the validity and robustness of the data collected.

The use of automated processes to collect data proved to be successful in attaining a wide range of results at a much quicker rate than manual operation. Scripts customized to specific environments provided the key for obtaining data; fortunately, the GCMD dataset records were uniformly structured but not all potential dataset sources will be as ordered, necessitating a more complex script to be written. Even with these processes in place to extract publication and citation data, a number of manual processes still had to be executed in order for the automated collection to occur, as seen with the attainment of secondary publication citation counts. In addition, the cleaning and processing of retrieved outputs was immensely time intensive but expected given the vast quantity imported.

Study limitations
There are several limitations with this study. The analyzed results lack generalizability to dataset reuse in earth science research since only ‘well-behaved’ or complete dataset records were considered in the sample for analysis. To a certain extent, these included a number of records from established data centers that have the knowledge and skill base to keep and curate datasets for long-term accessibility and use, and would therefore be expected to maintain more complete and well-documented metadata records. There were over 5,000 dataset records collected for this study that were not examined but would provide a needed complement to this analysis and could illuminate a dimension previously overlooked. In addition, different sampling techniques could have been used to procure the dataset records analyzed; the datasets selected were based on a convenience sample of the earth science domain.

The measure of disciplinary reach also warrants criticism as it currently only accounts for publications that were published in scholarly journals. As shown by the dataset metadata record, the definition of ‘publication’ covers a wide breadth of scholarly works that have also garnered citations. It is important to consider that dataset reuse may not always result in a journal article publication but may serve non-research functions such as the development of public policy or educational training (Fienberg, Martin, & Straf, 1985).

FUTURE DIRECTIONS AND IMPLICATIONS
These results only demonstrate a small portion of the prospective analysis on dataset reuse tracking in the scholarly literature. For this preliminary study, only “well-behaved” datasets were selected for analysis. The challenge will be to extend the techniques developed for dataset records that are not necessarily as complete.

In terms of using disciplinary reach as a measure for impact, a comparison study with datasets in other disciplines and research areas should be conducted to assess the robustness of this indicator. Performing a cluster analysis on publications and related journal subject classifications for each dataset is a proposed next step to understanding the relationship between the dataset and publication impact. A key component that would contribute to the significance of this measure is the temporal aspect attributed to the dataset and its public release, along with the affiliated publications and their circulation in the scholarly literature. Taking into account these factors may elucidate additional elements to consider for evaluating dataset reuse and its future influence on scholarship.

A number of new indicators and citation databases have emerged in the past two decades, providing new opportunities for bibliometric research (Leydesdorff, 2009). This evolving environment is also witnessing an increase in access to different types of scholarly literature beyond peer-reviewed journal articles, information resources that contribute just as much to advancing the collective knowledge base but have thus far remained relatively unrecognized. Research datasets are no exception and provide unique challenges for citation and the subsequent adoption of citation-based analytical techniques.

This preliminary study examined the impact of earth science datasets reuse in the literature and demonstrated that the reach and potential reuse of these datasets by scholars outside of the original field of creation (i.e. physical science) could be observed based on citations of publications in multidisciplinary journals. Within scientific research, the increase in accessibility to digital data has initiated new types of research that were otherwise not possible to conduct (Faniel & Zimmerman, 2011). Similar to a citation, a dataset can be used in multiple ways for multiple purposes but capturing the landscape of its use through relative impact offers potential incentives not only to the researcher and also influences policy and program development for long-term data management and use.

ACKNOWLEDGMENTS
The author would like to express her immense gratitude to Vetle I. Torvik for his guidance and profound insight on applying automated processes to large-scale datasets,
REFERENCES


Davis, H. M., & V.


Garfield, E. (1972). Citation analysis as a tool in journal evaluation – journals can be ranked by frequency and impact of citations for science policy studies. Science, (178), 471-9.


