8-2016

Acknowledgement Lag and Impact: Domain Differences in Published Research Supported by the National Science Foundation

Monica Inez Ihli
University of Tennessee, Knoxville, mihli1@utk.edu

Recommended Citation

Ihli, Monica Inez, "Acknowledgement Lag and Impact: Domain Differences in Published Research Supported by the National Science Foundation. " Master's Thesis, University of Tennessee, 2016.
https://trace.tennessee.edu/utk_gradthes/4045

To the Graduate Council:

I am submitting herewith a thesis written by Monica Inez Ihli entitled "Acknowledgement Lag and Impact: Domain Differences in Published Research Supported by the National Science Foundation." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Information Sciences.

Carol Tenopir, Major Professor

We have read this thesis and recommend its acceptance:

Suzanne L. Allard, Awa Zhu

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Acknowledgement Lag and Impact:
Domain Differences in Published Research Supported by the National Science Foundation

A Thesis Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Monica Inez Ihli
August 2016

Dedication

This thesis is dedicated to my father, Ralph E. Ihli, Sr., who passed away in February 2016. He was a man who embodied perseverance and compassion. In a way, I always knew that I chose this topic for him because we both loved science, and he would be pleased that I had written a thesis about science. When I set out to do this work, I often imagined the day that I was finished and could show him what I had accomplished. He had been the first person I called when I found out I had succeeded in obtaining an internship at NASA. He was the first person I called when I learned that my first peer-reviewed paper was going to be published. He was going to be the first person I called when I succeeded in defending my thesis.

Acknowledgements

I would like to express my sincere gratitude to the chair of my thesis committee, Dr. Carol Tenopir, whose support gave me the confidence to proceed with undertaking this project. I also thank Dr. Suzie Allard and Dr. Xiaohua Zhu, both of whom also served on my thesis committee. Dr. Zhu has been a close advisor since the beginning of my time in this program, frequently offering me encouragement and good advice. I am thankful to Dr. Allard for all the times she created opportunities for me to meet people and engage in the community of scholarship. These opportunities provided the inspiration from which this project originated.

Additionally, I thank Corey Halaychik of University of Tennessee Libraries as well as Timothy Otto from Thomson Reuters for their assistance in facilitating access to the premium Web of Science API service, which made this project possible.

Abstract

This research combined archives of grant awards with a five-year period of bibliographic data from Web of Science in order to conduct an input-output study of research supported by the National Science Foundation. Acknowledgement lag is proposed as a new bibliometric term, defined as the time elapsed between when a grant is awarded and when a document is published which acknowledges that award. Acknowledgement lag was computed for the dataset, and domain differences in lag times were analyzed. Some areas, such as Plant & Animal Science or Social Science, were found to be more likely than other categories to acknowledge a grant seven or more years later, while other categories, such as Physics, were most likely to publish a grant acknowledgement in two years or less. In addition, rank-normalized impact factors were computed for journals in which these articles were published, as a measure of journal impact that is comparable across categories of research. The overall distribution of rank-normalized journal impact factors for research articles acknowledging support by the National Science Foundation was analyzed. Category-level analysis was also performed, and it was found that there were differences in the journal impact factor trends for publications from different domains in the dataset. Research in Materials Science was substantially more likely than other categories to publish in the most elite journals of its respective domain. Social Sciences research was also found to be one of the strongest research areas in terms of impact factor, despite being one of the smaller categories in terms of publication counts. However, other categories were found to be disproportionately more likely to have been published in lower impact factor journals for their respective fields, such as Mathematics and Computer Science. The methodology developed in this project demonstrates a workflow that could be implemented by the NSF or other agencies. The findings demonstrate that systematically linking grants to publications can yield information of strategic value, allowing agencies to better understand field differences in outcomes and providing a means for tracking changes in publication-related metrics over time.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

IF	Impact factor
ISI	Institute for Scientific Information
JCR	Journal Citation Reports
NSF	National Science Foundation
rnIF	Rank-normalized impact factor
SCIE	Science Citation Index Expanded
SEI	Science and Engineering Indicators
SSCI	Social Science Citation Index
WoS	Web of Science

Chapter 1

Introduction and General Information

The National Science Foundation (NSF) was established in 1950 to promote science in the United States ("National Science Foundation Act of 1950," 1950). The NSF supports and monitors all fields of science and engineering, with the exclusion of medical sciences (National Science Foundation, n.d.). One of the functions of the NSF is to prepare the biannual publication of the Science and Engineering Indicators (SEI). The purpose of the report is to, “present information on science, technology, engineering, and mathematics education at all levels,” as well as research and development performance, public attitudes, and overall competitiveness of the United States. Indicators are described as “quantitative representations that might reasonably be though to provide summary information bearing on the scope, quality, and vitality of the science and engineering enterprise,” (National Science Board, 2014). The Science and Engineering Indicators serve as an example of how we think about and measure science in the United States of America. We count institutions and how much funding they receive. We count the publications, who publishes them, and what institutions those authors were affiliated with. We tend to describe research either in terms of the funding or the outputs, but seldom one in context of the other. Approaches to that effect appear to still be a developing area of metrics.

For the United States, one contributing factor for why it is uncommon to see large-scale analysis of research outcomes in the context of funding may be due to that researchers in this country experience a different funding environment than researchers in many other parts of the world. Jongbloed and Vossensteyn (2001) describe the United States' research funding approach to be inherently competitive and performance-oriented, observing that, "unlike their European counterparts, the American universities do not receive substantial amounts of funds as core funding for basic research," and that, "The universities in the US therefore have to compete for the bulk of their research funding, whereas many European universities often receive historically based allocations for research from their governments or funding councils." Where the amount of funding received in one year depends partially on performance outcomes from the previous year, it would make sense that methods for analyzing outcomes in the context of funding become increasingly important.

By contrast, institutional performance measures may be observed as a general indicator of success in the United States, but they are not systematically incorporated into the distribution of federal research funding. The 2014 Science & Engineering Indicators report likewise acknowledges that other countries tend to provide general university fund "block grants" to academic institutions, which may be used at the discretion of the institution for costs, including research costs, as they see fit. But the United States differs in, "preferring instead to support specific, separately budgeted R&D projects," in a system where competitive process of peer review manages the majority of federal R&D funding to academic institutions (National Science Board, 2014).

Although broad analysis of research outputs in the context of funding inputs may not be systematically employed in the United States at this time, there is certainly the potential and an interest to do so. The development of such methods relates closely relates to the field of bibliometrics. Bibliometrics is a domain of library science which employs statistical analysis upon documents to better understand trends within the documents, the use of documents, or changes within a body of literature over time (Broadus, 1987; Pritchard, 1969). Bibliometrics, although not without controversy, can help us to gain insights into how a field has developed over time, the contributions of authors, or the impacts of journals. Bibliometrics has an inherent usefulness to the study of award outcomes because it is expected that knowledge and findings from the supported research will be shared through publication in scholarly journals, among other forms of dissemination. However, it has historically been considered difficult to systematically describe and evaluate funding and research outcomes from an end-to-end perspective, largely due to that the information systems which track awards and the systems which track scholarly publications are typically maintained by separate entities. This contributes to the difficulty of conducting what Boyack and Jordan (2011) refer to as "input-output studies".

Systematic linking of grants and articles would have the potential to greatly enrich analysis and reporting of science, by enabling bibliometrics to be incorporated with other measures. For example, bibliometrics is often concerned with the time it takes for things to happen. There are several forms of lag that are established bibliometric measures: *citation lag*, which is a measure of the time between the publication of an article and the publication of a new article which cites the first; *indexing lag* which measures the time between when an article is published and when it is indexed; and *publishing* lag, which measures the time between when a manuscript is submitted or accepted by a refereed journal, and the when it is published (Diodato, 1994). If data about when grants were awarded and when articles were published could be systematically collected and linked, then it would be possible to calculate the difference as a form of lag. To that effect, the author of this study proposes the term **acknowledgement lag**, defined as the time elapsed between when a grant is awarded and when an article is published which acknowledges that grant. An additional point to consider for acknowledgement lag is that field differences are commonly looked for, and usually found, in bibliometrics. For example, citation practices, citation frequency, publication frequency, and field size vary from one domain to the next (Bornmann & Daniel, 2008), and studies have also found field differences for publication lag as well (Björk & Solomon, 2013). We might reasonably expect, therefore, that an analysis of acknowledgement lag might also reveal differences across fields.

Journal impact factor is a different example of an existing bibliometric approach which could be used to assess award outcomes in terms of publications. **Impact factor** (IF) is defined as the ratio of the number of citations in the current year for citable items within a journal to the number of source items published in the same journal over the last two years (Garfield, 1999). Journal impact factor is a metric which has been computed as part of the Institute for Scientific Information's (ISI) Journal Citation Reports since the mid-seventies (De Bellis, 2009, pp. 181-187). Impact factors are considered by many to serve as an indicator of influence. They may be used to evaluate journal subscriptions or as a contributing factor in evaluation of researchers. Assuming that it was possible to systematically link awards to published research, the distribution of impact factors for journals of publication is one example of a metric that could be considered.

However, impact factors vary across disciplines for reasons such as citation practices, publication and citation lag, and the indexing of journals which are being cited. Impact factors have also been found to have experienced inflation over time, due factors such as a trend towards increasing numbers of citations in reference lists (Althouse, West, Bergstrom, & Bergstrom, 2009). Because the ranges of journal impact factors vary across disciplines, it has been said that “all citation studies should be normalized to take into account variables such as field, or discipline, and citation practices,” (Garfield, 1999). One solution that has been offered to the problem of categorical differences is the conversion of impact factors to **rank-normalized impact factors** (rnIF), defined as the value derived from a rank-based normalization procedure which may be used to facilitate cross-category comparisons of impact factors. A journal’s rnIF is based on the journal’s position when all journals in a given category are ordered based on their respective impact factors (Pudovkin & Garfield, 2004).

It should be noted that the adoption of journal impact factors as a proxy for journal quality or success has experienced its fair share of criticism and controversy. The use of journal-level impact factors has been said to ignore differences in individual article citation rates, and it has been argued that impact factors are biased towards English language journals (Seglen, 1997). Others have argued that the use of impact factor as a measurement of prestige is biased against international journals from peripheral countries, whose journals may not be indexed by ISI (Bordons, Fernández, & Gómez, 2002). Still others are simply frustrated with the extent to which publication in journals with high impact factor has become a requisite for individual success, and call for a greater emphasis on qualitative evaluation of research over quantitative (Verma, 2015). Despite the controversy, the proponents of the journal impact factor are often as vocal as those who oppose such measures. In the past, it has been said that, “Impact factor is not a perfect tool to measure the quality of articles, but there is nothing better and it has the advantage of already being in existence and is, therefore, a good technique for scientific evaluation,” (Hoeffel, 1998). A study validating the IF as a proxy of citation frequency has supported the position that “blanket criticism of using the IF for decisions in research funding is therefore at least partially exaggerated,” (Racki, 2009). In an ideal world, we would always have qualitative knowledge of every article. However, that is seldom possible for large-scale reporting, and so summary and quantitative measures such as journal impact factor have their place. Readers desiring a more in-depth review of history and debate surrounding the impact factor are directed to Chapter 7 of De Bellis (2009). In recent years, there have been efforts to propose alternatives to journal impact factor (Haustein et al., 2014; Leydesdorff, 2012; Leydesdorff & Opthof, 2010; Piwowar, 2013), but it remains to be seen how widely adopted or institutionalized they will become by comparison.

Reliance on citation-based measures as an indicator of success is recognized in the SEI reports, in that publication counts and highly-cited articles for United States authors are reported and compared to other countries. But it is important to note these measures can be arrived at from bibliometric data alone, without considering specifically the relationship between the publications and United States agencies which may have provided financial support for that research. It would be useful to know the distribution of impact factor for journals of publications within an agency such as the NSF, or how impact may be distributed across the agency's research portfolio. Realistically, there are most likely efforts within individual programs and divisions of the NSF and other agencies to collection such data. But lacking systematic methods of data collection and linkage, however, this level of analysis would be difficult to achieve.

Linking awards to bibliographic data would enrich science reporting, and the landscape of systems and practices which might support such analysis continues to evolve as funding agencies, publishers, and bibliographic databases take steps in that direction. The National Science Foundation itself is working to increase availability and access to research supported by the agency, by stepping up requirements for investigators to report and make available resulting publications and data (https://www.nsf.gov/news/special_reports/public_access/). The NSF's Public Access repository has the potential to eventually become a reliable source of data for linking publications to grants. But as of the present, the data is sparse and the public access repository is described as being in beta. The repository website presently states of the new reporting and depository requirements that they, "will apply to new awards resulting from proposals submitted, or due, on or after the effective date of the Proposal & Award Policies & Procedures Guide (PAPPG) that will be issued in January 2016," suggesting that the availability data within the system can be expected to improve with time.

Although the NSF is still working with some publishers and other entities in the industry to develop its system, bibliographic databases have been working to bridge publications to funding sources for some time now. Thomson Reuters' Web of Science (WoS) is a bibliographic database service that began indexing funding acknowledgements and grant information (where reported), as early as August 2009 (Thomson Reuters, 2009). Subsequent research has sought to explore and define the completeness and reliability of this data. Tang, Hu, and Liu (2016) report limitations of WoS funding acknowledgement data to include that only acknowledgements in SCIE are systematically recorded, and that SSCI is underrepresented. In a sample of approximately 9.7 million SCIE records from 2009-2014, approximately 4.6 million contained funding acknowledgement data, while only about 250 thousand out of 1.5 million SSCI records for the same period contained acknowledgement data. Some of the complications of correctly identifying funding sources within acknowledgements can include that a source might be referenced in a variety of ways, such as sometimes using acronyms, using name variants, including only a grant number, or referencing a parent organization. Consequently, the process of retrieving articles from WoS based on funding acknowledgement or grant can be "hit or miss" (Coppin, 2013).

In summary, the indexing of funding acknowledgements by bibliographic databases has created new opportunities in the development of metrics and workflows for analyzing funded research. Despite that limitations clearly exist, grant acknowledgements within a publication can be used as a key to locate other information about the grant, such as when the grant was awarded. Identifying publications supported by an award facilitates the use of bibliometric measures such as journal impact factor to describe outcomes supported by the award. The value of the data to funding agencies would be increased if it were found to be suitable for automated analysis in a way that does not rely on manual intervention. This research attempts to explore the development of automated workflows for integrating award data with bibliographic data, as well demonstrating the usefulness of bibliometric measures derived from their integration. Furthermore, the author has for the first time proposed the term acknowledgement lag to describe a new bibliographic measure for the time elapsed between when an award has been granted and when an article was published which acknowledged the grant.

Chapter 2

Literature Review

Analysis of existing literature shows that there is still much to be done for exploring the analytical capabilities of linking awards to publications. Some studies evaluate whether or not funding and acknowledgement have a relationship to the impact or citations of a publication. Few of these are true input-output studies in the sense that they directly link grants to articles, and it is more common to consider publication acknowledgement at the level of the agency rather than at the level of the grant. Acknowledgement lag has not been found to be addressed in any studies which could be found. Impact factor is addressed in some of these studies, but they are not found to address categorical differences in research for an agency.

Analysis for a Specific Agency

Bibliometrics can be a useful tool for describing the bigger picture of the research being supported by specific agencies. For example, Belter (2013) analyzed 409 articles published between 2002 and April 2012 which acknowledged funding support from the Office of Ocean Exploration (OER) within the United States National Oceanic and Atmospheric Administration (NOAA). Bibliographic data for this study was identified through a combination of data internal to OER and searching Web of Science based on funding acknowledgement. This study did not directly link awards to publications. Belter concluded that the distribution of these publications was concentrated in certain regions of the United States, that overall article publication rates were variable over time, and that publication of research funded by NOAA OER tended to fall within several Web of

Science categories. Bibliometric analysis covered citation, authorship, and semantic aspects of publication data. Some examples of analysis include the of the number of articles published over time, institutional publication statistics and mapping, distribution for categories of publication, categorical distribution of citations to the articles, and percentile rank analysis.

Comparing Funded to Unfunded Publications

A different approach to incorporating funding information into bibliometrics may be found in studies which compare metrics for publications acknowledging funding support to those that do not. Some studies have investigated if acknowledgement of external research funding has any relationship to the quality or reception of published research. Boyack and Jordan (2011) analyzed over 1.4 million articles published between 1980 and 2009, and over 200,000 United States National Institutes of Health (NIH) funded grants. Grants and articles were directly linked in this case. Articles which included grant acknowledgements to either the NIH or the US Public Health Service (which includes NIH) were found to have been cited twice as often articles with no funding source identified. They also reported a variety of statistics about the dataset, including average number of articles per grant, average number of cites to articles per grant, and a time series analysis of grant-related quantities by initial grant year. Data for this study came from systems internal to the NIH which combine internal data with publication records from PubMed, while citation data was linked to from Scopus (another bibliographic database).

Zhao (2010) analyzed 266 articles published between 1998 and 2008 across seven journals in library and information science, determining that those which acknowledged grant funding were cited, on average, over 40% more often than articles which did not acknowledge funding. The Scopus bibliographic database was used for this analysis. Zhao also reported the distribution of citations per year for funded and non-funded research, as well as distributions for several other attributes, such as funding agency and institutional affiliation. Countering Zhao, however, Rigby (2011) used Web of Science to analyze 301 papers from the journal Cell and 3,414 papers from Physical Review Letters, and argued that any relationship between the number of funding organizations and the citation impact was weak at best. Rigby's position may be in the minority, however, as more research seems to support the idea that funding positively relates to citations than refute it.

Jowkar, Didegah, and Gazni (2011) analyzed bibliographic data for over 80,000 articles published by Iranian authors between 2000 and 2009. They were interested in determining the proportion of articles which were funded and whether or not being funded seemed to have any effect on the rate of citations, as well as looking at differences in subject areas. Data was extracted from Web of Science's Science Citation Index Expanded, as well as the companion product Conference Proceedings Citation Index. Publications were classified using the Essential Science Indicators schema (not to be confused with the NSF's report titled similarly) by Thomson Reuters (2015). It must be presumed that conference proceedings were not classified, as the ESI data is a mapping table for journals. They found only 12.5% of publications for their sample based on

Iranian authorship to acknowledge funding. However, they did find, similarly to other investigators, that funded research tended to produce more citations than unfunded research. This finding held true across most subject categories.

Wang and Shapira (2015) used a text-mining approach on a dataset of 89,000 bibliographic records for nanotechnology publications in order to identify papers with funding acknowledgements. Data was collected using Web of Science. The researchers found that papers with such acknowledgements were more likely to have been published in high impact journals, as well as being more likely to have received a higher number of citations. They also found that funding diversity in terms of international collaboration had some positive relationship to journal impact factor.

Incorporated Funding Amounts into Bibliometric Analysis.

Some studies have attempted to explore the role and effect of funding expenditures in relation to bibliometric outcomes. Auranen and Nieminen (2010) wanted to explore if national approaches to science funding policy, such as more competitive systems of distributing research allocations to universities in nations where universities receive annual allocations, are more efficient in producing scientific publications. This study is particularly interesting in that it deals directly with the previously referenced differences in funding environments described by Jongbloed and Vossensteyn (2001), and is a useful reference for any individual from the United States who wishes to better understand how other nations have historically approached funding allocation. The researcher’s analytical framework considers the overall mix of external versus internal or core funding for several countries as inputs, comparing these to publications as outputs.

A variety of national sources were used to collection expenditures data, and bibliographic data was taken from Web of Science databases. They calculated the ratio of research and development expenditures to publications as a measure of efficiency. Calculations were performed at the national level for Australia, Denmark, Finland, Germany, Netherlands, Norway, Sweden, and the United Kingdom. While the data did not support a straight-forward, cause-and-effect relationship between competitiveness of policy and efficiency, the researchers argue that some nations did experience a significant increase in efficiency over time while others remained relatively stable. As a final note, although this study did assess both funding expenditures and outcomes, these values were taken as aggregates of both without considering direct relationships between the former and the later. This was not an input-output study in the sense of directly linking grants to articles.

A different study (MacLean, Davies, Lewison, & Anderson, 1998) took the approach of analyzing the distribution of the number of funding agencies acknowledged in papers for research on malaria, finding the data to suggest that “the most highly cited papers acknowledge support from more funding bodies than papers with low citation scores, and papers with progressively more funding bodies have a higher impact.” The researchers could not, however, find evidence of a direct correlation between funding dollars as input and citations as output. For this study, data about funding sources was collected by asking organizations who sponsored malaria research for records of grants awarded. Bibliographic data was gathered by selectively retrieving records published in 1989 from the Science Citation Index based on keywords identified as topically relevant. The relationships directly linking grants to articles were established by manually examining the 776 articles, of which a funding organization was identified (either explicitly by acknowledgement or implicitly by author address) in 758 cases.

A recent report in Canada compared the amount of research funds awarded to researchers in Quebec, Canada over a period of fifteen years to research outcomes in terms of publications derived from Web of Science (Mongeon, Brodeur, Beaudry, & Larivière, 2015). They identified the number of researchers funded through several agencies and compared these to the number who did not receive funding, finding that “the number of publications is strongly linked to the amount of funding received by researchers.” However, it must be cautioned that in this case the unit being analyzed is the researcher and not the grant. Rather than linking awards to papers directly, their methodology was to estimate the funds received by each researcher based on agency records, and then count the number of papers that researcher had published over a certain period of time.

To summarize the literature, only two out of nine studies analyzed involved direct mapping of grants to articles. Typical questions asked of the data included considering how articles with funding acknowledgements compared to articles without funding acknowledgement, looking to see if funding influenced the number of publications, and characterizing the research supported by an agency using standard metrics such as citation and publication counts. Seven of the studies drew upon Web of Science citation indexes for bibliographic data, while one relied on Scopus alone and another combined data from both PubMed and Scopus. Analysis involving manual work and evaluation was used for smaller samples in the hundreds of records, but automated processing was necessary to realistically handle larger numbers of records in the thousands or hundreds of thousands.

Research Questions

The literature establishes that there is definite interest in relating grants and publications, but that there is still work to be done in exploring how this may be accomplished and what we can learn from it.

Automated processing tended to rely on the emergence of indexed funding acknowledgements and grants as a practice undertaken by bibliographic databases. Web of Science is not the only bibliographic database, but bibliometric researchers have invested in exploring the capabilities and limitations of WoS funding and grant data, and have published their findings to the benefit of others. Despite that we are thusly made aware that the data may be incomplete, particularly for Social Sciences, WoS is one of the best sources readily available at the present, and the literature establishes it to be an acceptable resource for this developing area of bibliometrics.

This research attempts to develop a workflow capable of demonstrating the value of integrating grant and publication data, by conducting a bibliometric analysis on publications which acknowledge support by the National Science Foundation. The NSF is an ideal organization for this kind of study, because of the agency's broad research portfolio covering most areas of science. This makes the data suitable for comparison across research domains. Two areas to be examined are the acknowledgement lag of research publications supported by the NSF, and the journal impact factor for publications acknowledging support by the NSF. The research questions for this study are as follows:

RQ1. What are the acknowledgement lag times between the award of a grant by the NSF and the publication of articles acknowledging the grant?

RQ2. Are there differences in acknowledgement lag for different categories of research which acknowledge support from the NSF?

RQ3. Regarding articles which acknowledge support by the NSF, what is the distribution of impact factor for journals of publication?

RQ4. Are there categorical differences in the distribution of journal impact factor for different fields of research?

Chapter 3

Data and Methods

Overview of Data Sources

Several things would be required of data for it to be capable of answering the research questions: First, the data should establish a link between grant awards and published articles as outcomes of the grant. Second, there must be some consistently-applied mechanism for identifying the impact factor of the journal of publication for these articles. Finally, the articles should be able to be systematically classified in order to look for differences in outcomes for categories of research supported by the NSF.

No single data source presently exists which could answer the research questions, but several data sources were integrated for this purpose. Figure 1 illustrates how data acquired from the funding agency was integrated with data from the Web of Science (WoS) bibliographic databases, as well as Journal Citation Reports and an Essential Science Indicators (ESI) categorization schema, in order to produce an enriched dataset which fulfilled these requirements. Each of these data sources and their characteristics will first be described individually. Following the overview of data sources, the processes by which data was extracted and integrated into a central project database will be reviewed in detail.

NSF awards.

Data describing grant awards was downloaded from the U.S. National Science Foundation’s online repository of awarded grants (https://www.nsf.gov/awardsearch/download.jsp). The NSF makes funding data

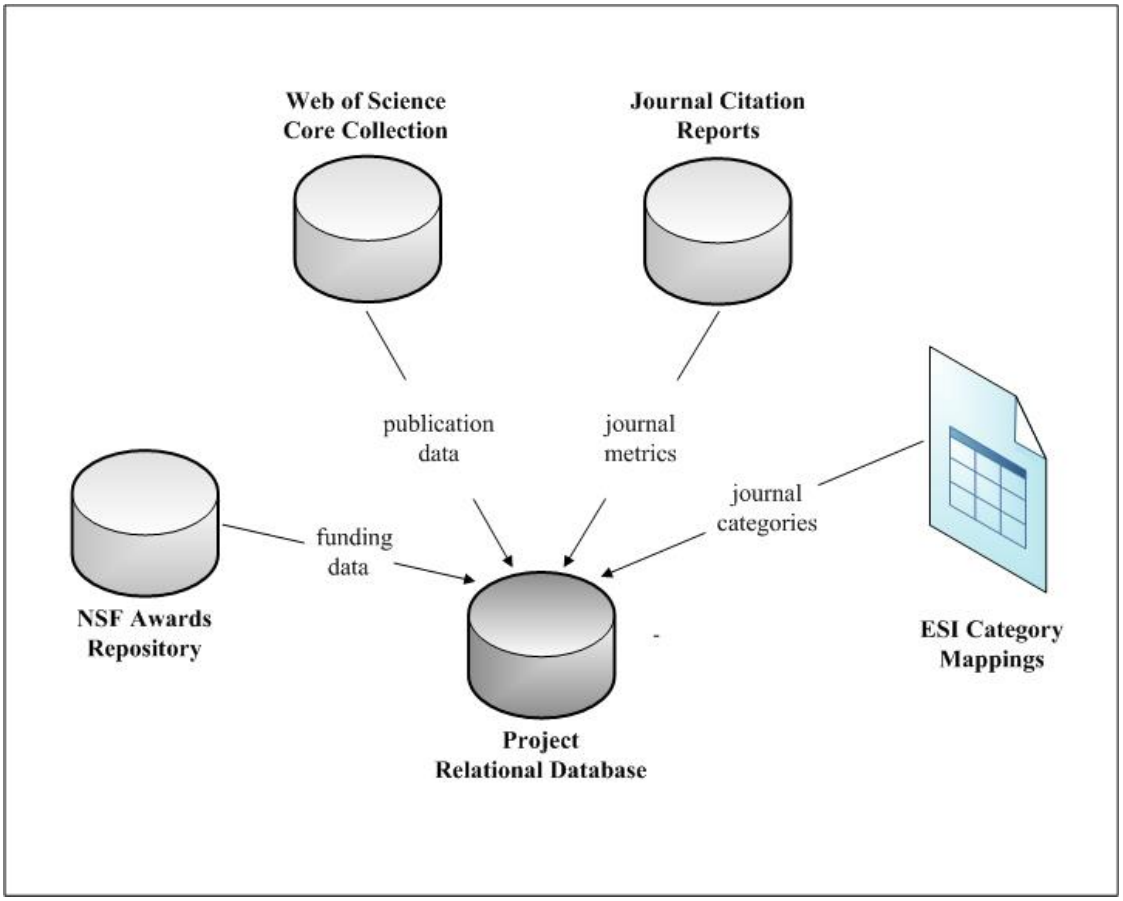

Figure 1. Data sources integrated for analysis in this project.

available in zip files organized by fiscal year. Each year's zip file contains a separate file for every grant that was awarded. Files are encoded in XML format. The XML schema is available online at https://www.nsf.gov/awardsearch/resources/Award.xsd. Data elements include basic information such as the grant award number, award amount, title, abstract, principle investigator (PI), and the PI's institution.

Article publication data.

Bibliographic data for articles was extracted from Thomson Reuters' Web of Science (WoS), accessed using the University of Tennessee's institutional subscription to the WoS Core Collection. The use of Web of Science for bibliographic data has the advantage of the data being maintained by the same entity responsible for both the Journal Citation Reports and the categorization schema which will next be described. This ensured that the names of journals were used consistently across data products—a requisite for linking journal data to impact factor and categories. At the time of this research, subscription to the Core Collection included:

- Science Citation Index Expanded (SCIE) (1900-present).
- Social Sciences Citation Index (SSCI) (1900-present).
- Arts & Humanities Citation Index (1975-present).
- Conference Proceedings Citation Index- Science (1990-present).
- Conference Proceedings Citation Index- Social Science & Humanities 1990-present).
- Book Citation Index– Science (2005-present).
- Book Citation Index– Social Sciences & Humanities (2005-present).

- Emerging Sources Citation Index (2015-present).
- Current Chemical Reactions (1985-present) (Includes Institut National de la Propriete Industrielle structure data back to 1840).
- Index Chemicus (1993-present).

Journal Citation Reports.

Journal Citations Reports (JCR) are available by subscription through the Incites interface by Thompson Reuters (http://jcr.incites.thomsonreuters.com/). Basic report data includes the full journal title, total citations to the journal for the given year, journal impact factor, and Eigenfactor score, although the inclusion of indicators is customizable and a variety of additional indicators are available. Data is available through this interface beginning with the year 1997. A review of past announcements regarding the availability of new reports would suggest that each year's JCR reports are made available around June of the following year. JCR report data has been made available as PDF, comma-separated value, or Excel spreadsheet file format. At the time of this research, the interface supported download of journal metrics, but it did not support categorization of the data as it was exported. Thus any categorization of journals within the JCR reports needed to be performed through additional processing. In addition to other optional journal-level bibliometric indicators, each row of data from the downloaded files contained the full journal title, global rank (across all categories), and impact factor for that journal, if available. Some newly indexed journals do not have an impact factor available for a given year, and in these cases the IF field for that row is populated with "Not Available".

Essential Science Indicators category mappings for journals.

In order to facilitate cross-categorical comparison, journals must be organized into subject areas. It has been proposed that a narrow subject classification schema is preferable for detailed analysis of small sets of publication data, whereas a broader schema is more suitable for general analysis across an organization or country (Thomson Reuters, 2015). Thomson Reuters does offer a classification schema described as the Web of Science (WoS) subject classification schema. This schema includes 232 subject categories. However, the categories are not mutually exclusive, and the number of categories is so large that it would be quite difficult to analyze differences using common statistical tests. By contrast, the Essential Science Indicators (ESI) schema is a broader classification schema including only 22 subject areas. The subject areas include science and social sciences, but exclude arts and humanities. Unlike the WoS schema, there is no overlap in category membership for the ESI schema. The broad scale of this research project, and the preference for mutually-exclusive category assignment, would suggest that the ESI category schema is the more appropriate selection. A third category of schema called the GIPP promises an even broader level of categorization (http://ipscience-help.thomsonreuters.com/inCites2Live/indicatorsGroup/aboutHandbook/appendix/mappingTable.html) with only six classes. However, it is indicated that there is significant overlap in journal categorization between the classes, and only six categories might not be granular enough for this analysis.

A version of the ESI schema mapping of journal titles to categories dated February 2016 was retrieved from http://ipscience-help.thomsonreuters.com

/incitesLiveESI/8289-TRS.html. Older versions of the mapping table may be found, such as a version dated 2012 which was found at http://ipscience-help.thomsonreuters.com/incitesLive/7622-TRS/version/default/part/AttachmentData/data/ESI_Journal_Category_Map_2012.xlsx, but there are occasional discrepancies between the newer and older versions. For example, *Advances in Artificial Intelligence* was classified into the Engineering category in the 2012 version, but Computer Science in the updated version. To avoid conflicting journal classifications, only the updated 2016 version of the schema was employed in this analysis. For the 2016 schema mappings, each row of data in the file included the full journal title, 29-character abbreviation of the title, 20-character abbreviation, ISSN, EISSN, and Category assignment for the journal.

Project Database

A MySQL database was used to store the data for the project. A high level representation of the design is represented in Figure 2. The ERD diagram shows how data from the various sources relate to one another. The diagram shows the final layout of tables and fields, although some fields (such as category or rank-normalized impact factor in the JCR table) were added in subsequent stages of processing after the data had been imported. Some grants were related by acknowledgement to one or more articles, but the nature of the data extraction process meant that an article would not be included in the ARTICLE table if it had not been found to acknowledge an award.

Most, but not all, articles were able to be mapped to a Journal Citation Report metric for the JCR year which corresponded to the publication year of the article,

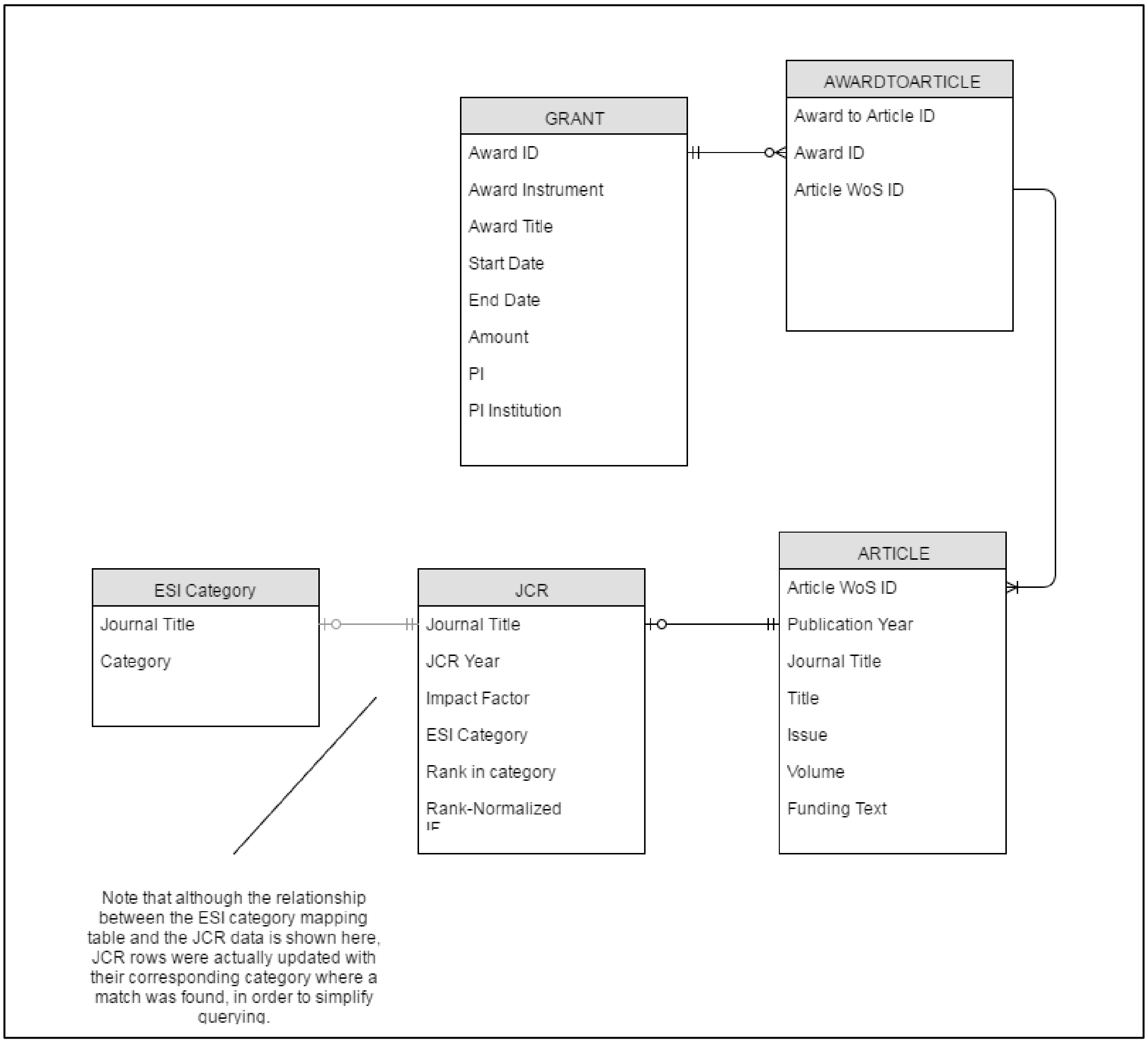

Figure 2. Entity Relationship Diagram for project database.

assuming that the journal title of the publication matched a journal entry in the Journal Citation Report. For the sake of showing how the ESI mapping table relates to other data in this approach, the ESI table is shown as an optional relationship to JCR data, although in practice the JCR rows were updated with their corresponding categories to simplify querying and other processing. Not every journal title in a JCR was able to be mapped to an ESI category. Additionally, it should be noted that in practice, each JCR year's data was imported as its own version of the JCR table, although the diagram in Figure 2 is simplified in that it shows only a single instance of the JCR table. This was due to that queries which required a join on two columns (journal and year) were rather slow by comparison to just joining articles for a specific publication year to a specific JCR table, using only the journal as the join column.

Grant information stored in the AWARD table and publication data stored in the ARTICLE table were related using an associative table, due to that one grant may be acknowledged in multiple publications, and a single publication may acknowledge more than one NSF grant.

Also, note that the NSF Award ID and Web of Science UID (a system number within the Web of Science database) are the true unique identifiers of award and article records, respectively. However, these fields were not enforced as primary keys when building the database, due to that the processes of extracting, transforming, and loading (ETL) data from different sources to combine for purposes of analysis were subject to certain considerations that would not apply to a transactional database. For example, the ETL process for extracting bibliographic records would inevitably extract multiple copies of the same record, in cases where more than one grant was acknowledged by this same publication. This would have violated the primary key constraint on the Web of Science UID, had it been enforced while performing ETL. Not doing so permitted the ETL scripts to proceed with their work, while additional post-extraction clean-up and quality analysis handled any issues of duplicate records and referential integrity.

Data Extraction, Loading, & Preprocessing

Extraction, preparation, & import of JCR data.

Journal Citation Reports data (including both Science Citation Index and Social Science Citation Index) were downloaded in comma-separated value file format from the Incites Journal Citation Reports interface for years 2010 through 2014, corresponding with the publication years for articles which were to be downloaded. For each year’s exported file, every row of data contained the full journal title, and impact factor for that journal, if available.

The process of preparing the JCR data files for import into the project database began with stripping unnecessary header and footer rows. All journal titles were converted into uppercase to avoid having to deal with case-inconsistencies later when matching JCR data to categories within the database. Removal of duplicate rows was also an important step at this stage, due to that several hundred duplicate rows may exist within the file. Any rows of journal data containing the value “Not Available” for impact factor were stripped, so as not to cause a type conflict when importing this field into a numeric datatype column in the database. Finally, each row of data was also coded with the JCR Year which the data represented. These edits were performed on the csv files in spreadsheet software prior to importing into the MySQL database. Table 1 summarizes the number of records imported into the database as a result of the JCR data preparation and import processes. Table 2 shows a sample of the JCR data as it was imported into the database.

Table 1. Summary of outcomes for JCR data preparation & import.

JCR Year	Rows Exported From InCites	Duplicates Removed	IF Not Available Removed	Total Rows of JCR data w/ IF Imported to MySQL Per Year
2010	10,804	492	56	10,256
2011	11,302	554	49	10,699
2012	11,518	582	38	10,898
2013	11,619	597	46	10,976
2014	11,813	613	39	11,161
				53,990

Categorization of JCR data.

At this point, each year's JCR data had been imported as a table. The JCR data tables were updated to add a column for Category, Rank, and rnIF (rank-normalized impact factor), as these values would be determined and added over the next several steps of processing.

The February 2016 version of the ESI mapping table was converted to a csv file and imported as a table into the database. Tables containing the imported JCR data were joined to the ESI Category mapping table using the full journal title as the join column.

Table 2. Sample of JCR data prepared for import into database.

Full Journal Title	**Journal IF**	**JCR Year**
CA-A CANCER JOURNAL FOR CLINICIANS	144.800	2014
NEW ENGLAND JOURNAL OF MEDICINE	55.873	2014
CHEMICAL REVIEWS	46.568	2014
LANCET	45.217	2014
NATURE REVIEWS DRUG DISCOVERY	41.908	2014
NATURE BIOTECHNOLOGY	41.514	2014
NATURE	41.456	2014
ANNUAL REVIEW OF IMMUNOLOGY	39.327	2014
NATURE REVIEWS MOLECULAR CELL BIOLOGY	37.806	2014

Where a match on journal title could be found, the Category column in the JCR table was updated with the value of the Category in the ESI mapping table. A sample of data illustrating the outcome of the join is shown in Table 3.

Table 3. Example of journal mapping outcome.

Full Journal Title	Journal IF	JCR Year	Category Name	Rank In Category	rnIF
CA-A CANCER JOURNAL FOR CLINICIANS	144.800	2014	CLINICAL MEDICINE		
NEW ENGLAND JOURNAL OF MEDICINE	55.873	2014	CLINICAL MEDICINE		
CHEMICAL REVIEWS	46.568	2014	CHEMISTRY		

The outcomes of the mapping process for each year's JCR data is described in Table 4, where the distribution of journal mappings across the 22 ESI categories is shown for each JCR year. The number of mappings for a given category in a given year also serves as the class size for computing the rank-normalized Impact Factor in the next stage.

The categorized contents of each JCR year's data table was next exported back out of the database for further processing, with the records being sorted by category and then impact factor. A PHP script (see Appendix A) handled computation of rank and rank-normalized impact factor. Rank-normalized impact factors were computed for each category and for every year. The equation used to categorically normalize impact factors

Table 4. Summary of record outcomes for mapping JCR data to ESI categories

Category	2010	2011	2012	2013	2014
AGRICULTURAL SCIENCES	283	300	308	313	320
BIOLOGY & BIOCHEMISTRY	351	361	373	387	404
CHEMISTRY	465	473	485	491	508
CLINICAL MEDICINE	1,562	1,649	1,706	1,756	1,815
COMPUTER SCIENCE	321	339	344	358	365
ECONOMICS & BUSINESS	464	504	518	526	532
ENGINEERING	726	750	767	787	812
ENVIRONMENT/ECOLOGY	264	280	295	305	318
GEOSCIENCES	355	364	374	382	385
IMMUNOLOGY	132	140	144	146	153
MATERIALS SCIENCE	285	294	310	319	332
MATHEMATICS	423	442	453	465	480
MICROBIOLOGY	94	101	105	110	112
MOLECULAR BIOLOGY & GENETICS	250	264	277	284	289
MULTIDISCIPLINARY	29	30	32	33	39
NEUROSCIENCE & BEHAVIOR	291	305	315	317	326
PHARMACOLOGY & TOXICOLOGY	235	246	250	255	261
PHYSICS	277	283	288	292	296
PLANT & ANIMAL SCIENCE	670	691	715	724	744
PSYCHIATRY/PSYCHOLOGY	533	554	571	579	602
SOCIAL SCIENCES, GENERAL	1,517	1,679	1,759	1,806	1,871
SPACE SCIENCE	48	49	49	52	53
Total Journals w/ IF Mapped to Category:	9,575	10,098	10,438	10,687	11,017
Number of Journals which Failed to Map:	681	601	460	289	144
Total JCR Journals Processed:	10,256	10,699	10,898	10,976	11,161

was the method proposed by Pudovkin and Garfield (2004), but substituting ESI categories. Any journals which had been unable to be categorically mapped were excluded from the computation of rnIF or any other further processing.

The exact steps for computing rnIF were as follows: First, the JCR data for a given year was grouped by category. The journals within the category were sorted in descending order according to impact factor, and each journal's position was coded as that journal's rank within its respective category. Following the equation:

$$rnIF_j = \frac{K - R_j + 1}{K}$$

Equation 1. Rank-Normalized Impact Factor

Rank normalized impact factor *$rnIF_j$* was computed where *R_j* is the rank (position) of journal *j* when all journals are sorted in descending order by impact factor, and *K* is the number of journals within the category. The result was a measure comparable across categories, such that the highest ranking journal within each category would have a *rnIF* of 1.0 while median journals would be near 0.5.

Following the example of Pudovkin and Garfield (2004), an application of this equation is here demonstrated using the case of the Agricultural Sciences category, to which 320 journals from JCR 2014 mapped, and in which the journal *Advances in Agronomy* ranked 16th out of the 320. Demonstrating the equation shown above, the values are as follows:

$$rnIF_{Advances\ in\ Agronomy} = \frac{320 - 16 + 1}{320} = 0.953$$

Although it has been suggested by that a journal's rnIF is mostly stable over time, the limitations of this assumption have not been established. Thus, the decision was made to calculate the rnIF for journals across each JCR year so that even all differences over the period of time for the analysis could be accounted for.

Once the PHP script had assigned a rank and computed rnIF for each journal in the JCR dataset, it updated the corresponding record in the JCR table in database with those values. An example of the resulting rows of data once the rnIF had been calculated and added is illustrated in Table 5.

Extraction of awards and bibliographic data.

PHP scripts handled most processes for extracting, transforming, and loading the data, which may be reviewed in Appendix B. The first step for preparing the data was to download each year's awards from the NSF awards repository. The script first cycled through the award files for each fiscal year, and inserted each award as a record into the AWARD table of the MySQL database. For every award examined, the script next made a call to the Web of Science API, requesting all bibliographic records which met the following conditions: The document type must be an article (as other types of records such as books or conference proceedings could not be evaluated for impact factor), the text of the funding organization field must contain "NSF" or "National Science Foundation", the text of the grant information field must contain the award number

Table 5. Sample of JCR Data with rnIF Computed

Full Journal Title	Journal IF	JCR Year	Category Name	Rank In Category	rnIF
AMERICAN LABORATORY	0.092	2014	CHEMISTRY	507	0.004
AFINIDAD	0.075	2014	CHEMISTRY	508	0.002
CA-A CANCER JOURNAL FOR CLINICIANS	144.8	2014	CLINICAL MEDICINE	1	1.000
NEW ENGLAND JOURNAL OF MEDICINE	55.873	2014	CLINICAL MEDICINE	2	0.999
LANCET	45.217	2014	CLINICAL MEDICINE	3	0.999

currently being examined, the article must have been published within the period of time being examined, from 2010 to 2014, and it must have been published after the award year. For each bibliographic record returned as a result, the article record was inserted into the ARTICLE table, and an association record was entered in the AWARDTOARTICLE associative table. This process began with the year 2014 and worked backwards through prior years. The extraction process was repeated for as many award years as continued to return a meaningful number of articles. By award year 1979, only a few articles had been returned for several years in a row, so the decision was made to stop collecting data at that point. This brought the total number of award years examined to 35, in terms of finding awards which had been acknowledged by journal articles published between 2010 and 2014. Figure 3 describes how many articles published between 2010 and 2014 could be matched to award years in a given year. Out of the 363,729 awards examined between FY 1979 and FY 2014, a total of 13,918 awards could be matched to one or more articles published between 2010 and 2014. The number of awards with a publication acknowledgement for this period peaks for award year 2009, with a long tail being seen for award years older than the mid-nineties.

The first preprocessing step for bibliographic data was deduplication of article records. The nature of the data extraction script meant that a duplicate article record would have been retrieved and inserted in any situation where a single article had acknowledged multiple NSF grants. An additional preprocessing step included to replace html character references with the appropriate character across titles. This was due that titles in records received from WoS sometimes included ‘&amp’ instead of ‘&’, for example, which would interfere with using titles as a match point to JCR.

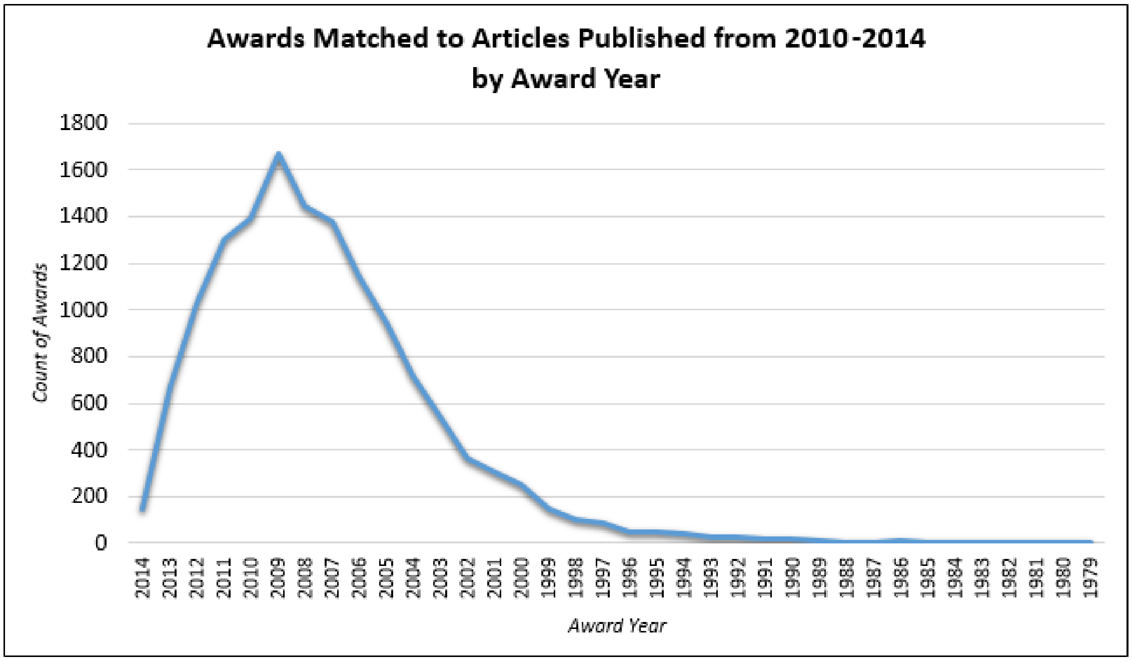

Figure 3. Awards matched to articles.

Table 6 summarizes the final number of unique articles extracted by publication year as well as the number of their acknowledgement relationships with grants. There were 58,495 articles extracted. A total of 66,740 relationships were formed between the 58,495 articles and the 13,918 awards. That a single article may acknowledge multiple grants explains why the count of relationships is greater than the number of articles.

Mapping of bibliographic data to JCR & categories.

Articles were mapped to JCR journal records corresponding to the publication year of the journal, and thus mapped to both categories and the rnIF of their respective

Table 6. Count of articles and their relationships with awards.

Publication Year	Articles	Article Relationships To Grants
2010	10,408	11,781
2011	11,418	12,998
2012	11,976	13,745
2013	12,342	14,142
2014	12,351	14,074
	58,495	66,740

journals during the year of publication as well. This join was again performed using the full journal title as the join column. 2,562 of the award-to-article relationships failed to map to a JCR record. Neither the category nor the rnIF of the journal of publication for the article could be identified for unmapped records, so the 2,562 were excluded from further analysis. This left 64,178 records to be evaluated.

Calculating lag between awards & articles.

Lag was computed as the difference between the award year and the publication year of the article, for every instance of an article acknowledging an award. The outcome is demonstrated in Table 7.

Methods of Analysis

To evaluate award output lag, a frequency distribution of lag values was computed within and across journal categories. Normality tests showed that the lag values were not normally distributed (p<.001). Therefore, a chi-square test of independence was selected as the appropriate non-parametric test to evaluate if lag was independent of category. To simplify interpretation of results, lag values for all observations were collapsed into four ranges: 0-2 years, 3-4 years, 5-6 years, and 7 or more years.

To evaluate impact, journal rnIF scores were likewise analyzed as a frequency distribution within and across categories. Normality tests showed that rnIF values were also not normally distributed (p<.001). Therefore, a chi-square test of independence was also selected to determine if the journal normalized impact factor scores for articles differed across categories. The rnIF observations were collapsed into the following ranges: .600 or less, .601 through .800, .801 through .900, and .901 or higher.

Table 7. Sample data for categories and lag computation

Award ID	Award Year	Article Title	Pub Year	Journal	rnIF	Category	Lag
1126860	2011	Characterization of indentation size effects in epoxy	2014	POLYMER TESTING	0.762	MATERIALS SCIENCE	3
1126862	2011	Three-dimensional flow measurements on flapping wings using synthetic aperture PIV	2014	EXPERIMENTS IN FLUIDS	0.75	ENGINEERING	3
1200011	2012	Simulated Adhesion between Realistic Hydrocarbon Materials: Effects of Composition, Roughness, and Contact Point	2014	LANGMUIR	0.874	CHEMISTRY	2
1200011	2012	Atomic-Scale Wear of Amorphous Hydrogenated Carbon during Intermittent Contact: A Combined Study Using Experiment, Simulation, and Theory	2014	ACS NANO	0.982	CHEMISTRY	2

Chapter 4

Results and Discussion

Acknowledgement Lag

RQ1 asked what are the acknowledgement lag times between the award of a grant by the NSF and the publication of articles acknowledging the grant. The frequency distribution of acknowledgement lag values for the 64,178 observations analyzed is shown in Figure 4. When plotted, the data formed a non-parametric, left-skewed curve with a long tail. The highest count of observances occurred at the 3-year mark. Overall frequency across lag ranges is summarized in Table 8. Approximately 63% of all observances showed a difference of 4 years or less between the time a grant was awarded and the time it was acknowledged in a publication. A difference of 5-6 years was found in 20.5% of all cases, while 16.6% of cases observed a difference of 7 or more years.

Table 8. Lag Distribution Across Lag Ranges

Lag Years	Frequency	Percent	Cumulative Percent
2 or less	18,287	28.5	28.5
3-4	22,038	34.3	62.8
5-6	13,171	20.5	83.4
7 or more	10,682	16.6	100.0
Total	64,178	100.0	

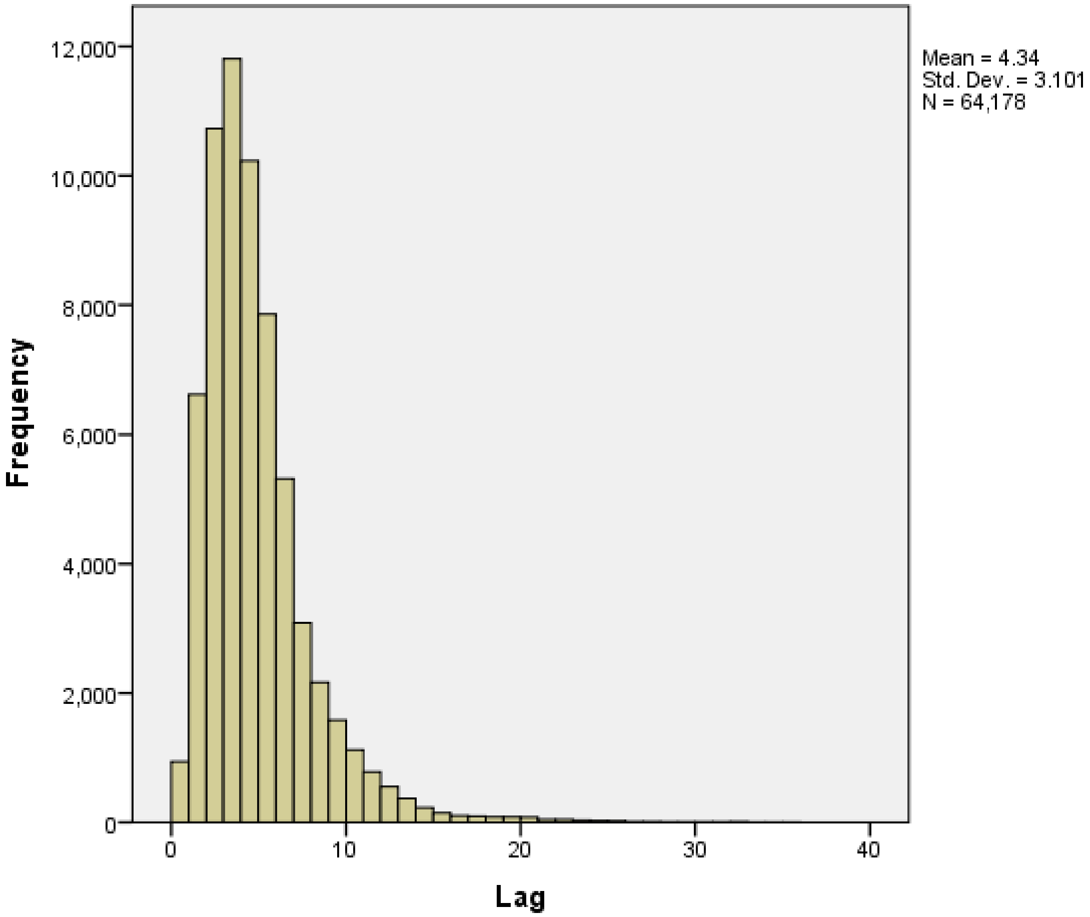

Figure 4. Frequency of lag observations.

RQ2 asked if there are differences in acknowledgement lag for different categories of research. To answer this question, more detailed descriptive statistics are shown for the same data across journal categories in Table 9. The median is a somewhat better measure of central tendency, given that we can see the effect of the handful of outliers from the long tail with very long lag times as they influence the mean for some categories, such as is the case for Environment/Ecology (the maximum value of 35 reveals that this category contains at least one such value). Categories with the highest median lag were Plant & Animal Science and Social Sciences at 6 years. At 5 years, Agricultural Sciences, Environment/Ecology, Geosciences, Immunology, Microbiology, and Psychiatry/Psychology also tended to show longer differences in years between award and publication, although in some cases there are so few observations (such as only 58 for Immunology) that any meaning interpreted from these results must be approached with caution. Categories with the shortest median difference between award and publication acknowledgement included Chemistry, Mathematics, Physics, and Space Science with a median of 3 years.

For acknowledgement lag values, a chi-square test of independence was run to determine whether or not differences between categories were statistically significant, with the results that X^2 = 6330.477, df=63, p < .001, which is significant and indicates that the distribution of lag does indeed differ by categories. The cross-tabulation across categories and across lag groups is shown in Table 10. With so many different combinations, it would be both tedious and unnecessary to compare every category to every other category, given that not every category features a meaningful discrepancy

Table 9. Descriptive Statistics for Lag by Category

Category	N	Mean	Median	Min	Max
AGRICULTURAL SCIENCES	201	5.90	5.00	1	20
BIOLOGY & BIOCHEMISTRY	2,379	4.75	4.00	0	31
CHEMISTRY	9,065	3.99	3.00	0	35
CLINICAL MEDICINE	370	4.13	4.00	0	13
COMPUTER SCIENCE	4,173	4.06	4.00	0	22
ECONOMICS & BUSINESS	201	4.31	4.00	0	13
ENGINEERING	4,264	4.26	4.00	0	25
ENVIRONMENT/ECOLOGY	3,107	6.19	5.00	0	33
GEOSCIENCES	4,178	5.66	5.00	0	32
IMMUNOLOGY	59	5.80	5.00	0	19
MATERIALS SCIENCE	3,390	4.17	4.00	0	32
MATHEMATICS	7,026	3.73	3.00	0	25
MICROBIOLOGY	546	5.66	5.00	0	27
MOLECULAR BIOLOGY & GENETICS	1,231	5.06	4.00	0	24
Multidisciplinary	2,445	4.54	4.00	0	26
NEUROSCIENCE & BEHAVIOR	596	4.75	4.00	0	24
PHARMACOLOGY & TOXICOLOGY	204	5.10	4.00	0	22
PHYSICS	11,789	3.35	3.00	0	24
PLANT & ANIMAL SCIENCE	3,642	6.39	6.00	0	34
PSYCHIATRY/PSYCHOLOGY	138	5.38	5.00	1	19
SOCIAL SCIENCES, GENERAL	504	6.41	6.00	0	29
SPACE SCIENCE	4,670	3.78	3.00	0	29

between actual and expected values. However, certain differences do stand out.

One of the things we may look for is to consider the adjusted residuals, which is a measure of the difference between observed and expected values for a cell in the cross-tabulation. The greater the adjusted residual, the greater that cell's contribution to the chi-square value indicated that differences exist between categories. The general rule of thumb is that an adjusted residual of +/- 2 indicates a discrepancy of interest, although this threshold may be increased to 3 or more when there are many cells. Adjusted residuals greater than +/- 3 have been identified with bold text in Table 10. To both better understand the discrepancies, and to generally understand the behavior of data within categories, we may consider distribution across lag classes for each category, and then see how different categories compare in terms of these distributions.

Attention is immediately called to Plant & Animal Science as containing the highest adjusted residual at 38.7 for the lag group of 7 or more years. Upon closer examination we see that, despite only 16.6% of overall lag observations falling in the range of 7 or more years, the distribution is much greater for Plant & Animal Science at 39.8% of observations within this category. Conversely, only 11.9% of lag observances for Plant & Animal Science fall within the range of 2 or less years. This suggests that researchers in Plant & Animal Sciences are more likely than those in other categories to publish an article acknowledging a grant 7 or more years after receiving the award, and are less likely than other categories of researchers to publish research acknowledging a grant within 2 or less years. Other categories which follow this same pattern of being less likely to publish a grant acknowledgement in 2 or less years and more likely to do so in 7

Table 10. Cross-Tabulation of Lag by Category

Category		**2 or Less Years**	**3-4 Years**	**5-6 Years**	**7 or More Years**	**Total**
Physics	Count	4,780	4,199	1,954	856	11,789
	% within Category	40.5%	35.6%	16.6%	7.3%	100.0%
	% within All	26.1%	19.1%	14.8%	8.0%	18.4%
	Adjusted Residual	**32.1**	3.2	**-11.7**	**-30.3**	
Chemistry	Count	3,127	3,096	1,617	1,225	9,065
	% within Category	34.5%	34.2%	17.8%	13.5%	100.0%
	% within All	17.1%	14.0%	12.3%	11.5%	14.1%
	Adjusted Residual	**13.7**	-.4	**-6.8**	**-8.6**	
Space Science	Count	1,438	1,870	914	448	4,670
	% within Category	30.8%	40.0%	19.6%	9.6%	100.0%
	% within All	7.9%	8.5%	6.9%	4.2%	7.3%
	Adjusted Residual	**3.6**	**8.5**	-1.7	**-13.4**	
Materials Science	Count	983	1,154	805	448	3,390
	% within Category	29.0%	34.0%	23.7%	13.2%	100.0%
	% within All	5.4%	5.2%	6.1%	4.2%	5.3%
	Adjusted Residual	.7	-.4	**4.8**	**-5.5**	
Clinical Medicine	Count	105	127	82	56	370
	% within Category	28.4%	34.3%	22.2%	15.1%	100.0%
	% within All	.6%	.6%	.6%	.5%	.6%
	Adjusted Residual	.0	.0	.8	-.8	
Multidisciplinary	Count	693	774	529	449	2,445
	% within Category	28.3%	31.7%	21.6%	18.4%	100.0%
	% within All	3.8%	3.5%	4.0%	4.2%	3.8%
	Adjusted Residual	-.2	-2.8	1.4	2.3	

Table 10 (Continued)

Category		2 or Less Years	3-4 Years	5-6 Years	7 or More Years	Total
Mathematics	Count	1,992	3,013	1,455	566	7,026
	% within Category	28.4%	42.9%	20.7%	8.1%	100.0%
	% within All	10.9%	13.7%	11.0%	5.3%	10.9%
	Adjusted Residual	-.3	**16.0**	.4	**-20.5**	
Economics & Business	Count	50	66	51	34	201
	% within Category	24.9%	32.8%	25.4%	16.9%	100.0%
	% within All	.3%	.3%	.4%	.3%	.3%
	Adjusted Residual	-1.1	-.4	1.7	.1	
Immunology	Count	10	16	13	20	59
	% within Category	16.9%	27.1%	22.0%	33.9%	100.0%
	% within All	.1%	.1%	.1%	.2%	.1%
	Adjusted Residual	-2.0	-1.2	.3	**3.6**	
Engineering	Count	1,156	1,491	961	656	4,264
	% within Category	27.1%	35.0%	22.5%	15.4%	100.0%
	% within All	6.3%	6.8%	7.3%	6.1%	6.6%
	Adjusted Residual	-2.1	.9	**3.4**	-2.3	
Computer Science	Count	1,126	1,563	885	599	4,173
	% within Category	27.0%	37.5%	21.2%	14.4%	100.0%
	% within All	6.2%	7.1%	6.7%	5.6%	6.5%
	Adjusted Residual	-2.2	**4.4**	1.1	**-4.1**	
Psychiatry / Psychology	Count	27	33	40	38	138
	% within Category	19.6%	23.9%	29.0%	27.5%	100.0%
	% within All	.1%	.1%	.3%	.4%	.2%
	Adjusted Residual	-2.3	-2.6	2.5	**3.4**	

Table 10 (Continued)

Category		**2 or Less Years**	**3-4 Years**	**5-6 Years**	**7 or More Years**	**Total**
Pharmacology & Toxicology	Count	39	66	47	52	204
	% within Category	19.1%	32.4%	23.0%	25.5%	100.0%
	% within All	.2%	.3%	.4%	.5%	.3%
	Adjusted Residual	**-3.0**	-.6	.9	**3.4**	
Neuroscience & Behavior	Count	134	203	144	115	596
	% within Category	22.5%	34.1%	24.2%	19.3%	100.0%
	% within All	.7%	.9%	1.1%	1.1%	.9%
	Adjusted Residual	**-3.3**	-.1	2.2	1.7	
Agricultural Sciences	Count	30	54	41	76	201
	% within Category	14.9%	26.9%	20.4%	37.8%	100.0%
	% within All	.2%	.2%	.3%	.7%	.3%
	Adjusted Residual	**-4.3**	-2.2	.0	**8.1**	
Biology & Biochemistry	Count	564	797	553	465	2,379
	% within Category	23.7%	33.5%	23.2%	19.5%	100.0%
	% within All	3.1%	3.6%	4.2%	4.4%	3.7%
	Adjusted Residual	**-5.3**	-.9	**3.4**	**3.9**	
Microbiology	Count	92	146	151	157	546
	% within Category	16.8%	26.7%	27.7%	28.8%	100.0%
	% within All	.5%	.7%	1.1%	1.5%	.9%
	Adjusted Residual	**-6.1**	**-3.8**	**4.1**	**7.6**	
Molecular Biology & Genetics	Count	252	379	275	325	1,231
	% within Category	20.5%	30.8%	22.3%	26.4%	100.0%
	% within All	1.4%	1.7%	2.1%	3.0%	1.9%
	Adjusted Residual	**-6.3**	-2.6	1.6	**9.3**	

Table 10 (Continued)

Category		**2 or Less Years**	**3-4 Years**	**5-6 Years**	**7 or More Years**	**Total**
Social Sciences, General	Count	69	114	128	193	504
	% within Category	13.7%	22.6%	25.4%	38.3%	100.0%
	% within All	.4%	.5%	1.0%	1.8%	.8%
	Adjusted Residual	**-7.4**	**-5.6**	2.7	**13.1**	
Geosciences	Count	757	1,163	942	1316	4,178
	% within Category	18.1%	27.8%	22.5%	31.5%	100.0%
	% within All	4.1%	5.3%	7.2%	12.3%	6.5%
	Adjusted Residual	**-15.4**	**-9.2**	**3.4**	**26.7**	
Environment / Ecology	Count	431	778	760	1,138	3,107
	% within Category	13.9%	25.0%	24.5%	36.6%	100.0%
	% within All	2.4%	3.5%	5.8%	10.7%	4.8%
	Adjusted Residual	**-18.5**	**-11.2**	**5.6**	**30.7**	
Plant & Animal Science	Count	432	936	824	1,450	3,642
	% within Category	11.9%	25.7%	22.6%	39.8%	100.0%
	% within All	2.4%	4.2%	6.3%	13.6%	5.7%
	Adjusted Residual	**-22.9**	**-11.3**	3.2	**38.7**	
Total	Count	18,287	22,038	13,171	10,682	64,178
	% within Category	28.5%	34.3%	20.5%	16.6%	100.0%
	% within All	100.0%	100.0%	100.0%	100.0%	100.0%

or more years include Environment/Ecology, Geosciences, Social Sciences, Molecular Biology & Genetics, Microbiology, Biology & Biochemistry, and Agricultural Sciences.

In contrast to these categories, lag values in Physics are observed to be proportionally higher than other categories for the 0-2 years of lag range. The distribution for this category decreases as the length in years of the lag range increases: 40.5% for 0-2 years, 35.6% for 3-4 years, 16.6% for 5-6 years, and 7.3% for 7 or more years. Researchers in Physics are more likely than researchers in other categories to publish a journal article acknowledging a grant within 2 years of receiving the grant, whereas they are less likely than other researchers to do so 5 years or beyond after receiving the grant. The data for Chemistry follow the same trend as the data for Physics.

Some categories are interesting, not simply because of their behavior at the extremes of publishing with an acknowledgement very quickly or very slowly compared to others, but because they have a higher proportion of observances occurring in the middle ranges of 3-4 years or 5-6 years. For 13 out of 22 categories, between 49% and 55% of observances are found in those middle lag ranges, and most of their cases the remaining observances dominate either one end or the other of the lag time spectrum. But in a few cases, we see an even higher number of observances concentrated in the middle ranges. In the case of Mathematics, for example, 63.6% of all observations fall in the ranges between 3 and 6 years, with 42.9% of that falling in the 3-4-year range. Mathematics is much less likely than other categories to publish in the 7+ years range, due to that most of their publications are happening from 3 to 4 years after the award. Space Science, while also being less likely than other categories to publish 7 or or more years after receiving the grant, peaks at 40% distribution for the 3-4 years range.

Research Impact

RQ3 asked what is the distribution of impact factor of journal of publication for articles which acknowledge financial support from the NSF. The distribution of rank-normalized impact factor scores for articles based on journal of publication is shown in Figure 5. The observed values form a right-skewed non-parametric curve. The distribution across collapsed ranges of rnIF are shown in Table 11.

Table 11. rnIF Distribution by Range

rnIF	Frequency	Percent	Cumulative Percent
.600 or less	10,550	16.4	16.4
.601 to .800	13,352	20.8	37.2
.801 to .900	17,634	27.5	64.7
.901 or higher	22,642	35.3	100.0
Total	64,178	100.0	

Recalling that the highest ranking journal within a journal’s respective category would have a rnIF of 1.0, while median journals would be near 0.5, it is interesting to observe that only 16.4 percent of all articles were published in journals whose rnIF were

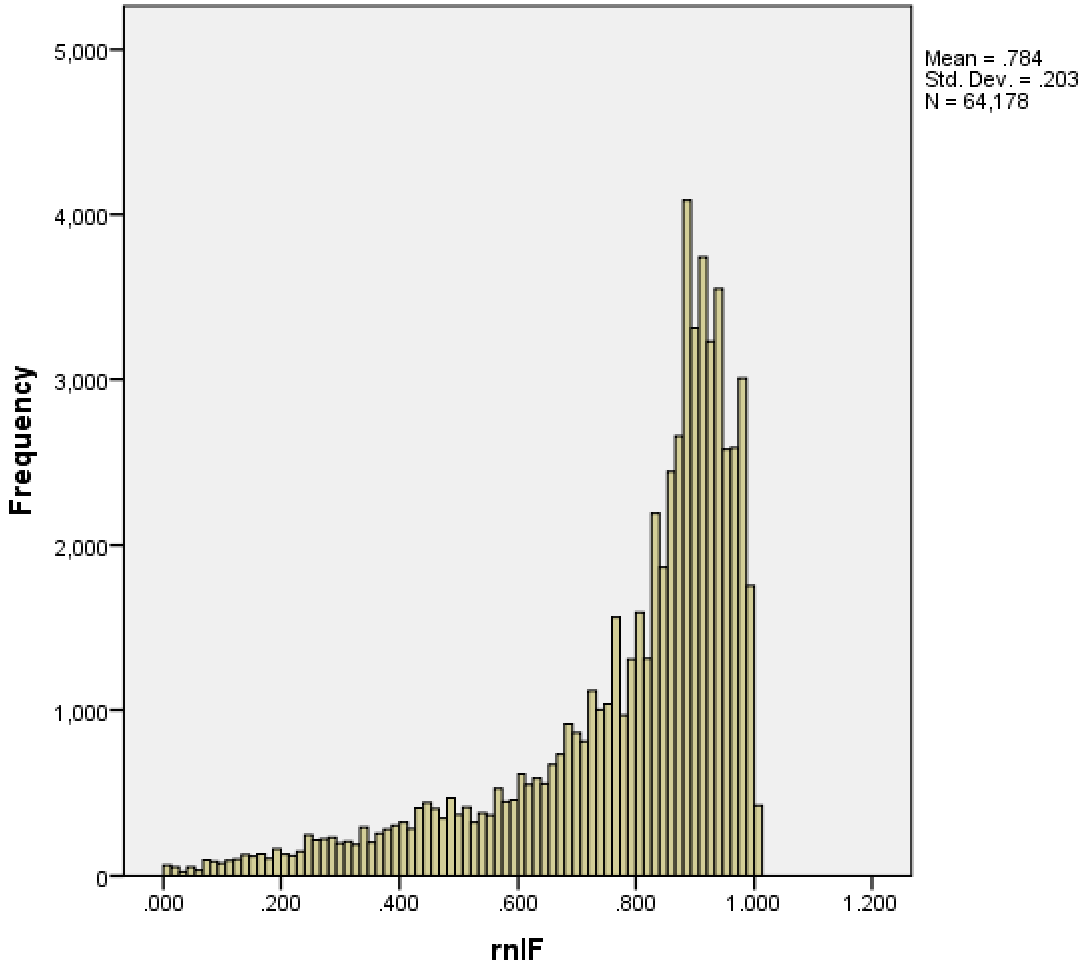

Figure 5. Frequency Distribution for rnIF

.600 or less, while 20.8 percent fell in the .601 to .800 range, 27.5 percent in the .801 to .900 range, and 35.3 percent fell in the .901 or higher range.

RQ4 asked if there are categorical differences in the distribution of journal impact factors for different fields of research. As noted in the methods, rnIF values were grouped into ranges so that a chi square test of independence could be computed, with the results that X^2 = 10090.609, df=63, p < .001, which is significant and is indicative of differences in rnIF ranges across categories. The cross-tabulation of categories and rnIF ranges is shown in Table 12, with the publication categories presented in order of greatest to least adjusted residual based on the .901 or higher rnIF range. The results show that the distributions within categories are typically not consistent with the distribution of the overall data. When interpreting the categorical distributions, recall that the adjusted residuals give a sense of how close or far the observed values are from the expected values for a publication category within a given rnIF range, and that these should be taken into account when looking at distributions within a category. The categories with the highest in-category distributions of articles in the top tier of journals are Materials Science (55.8%), Multidisciplinary (58.7%), and Social Sciences (59.7%), in addition to Chemistry (42.3%), Plant & Animal Science (45.7%), Geosciences (44.4%), Space Science (43.4%), and Environment/Ecology (41.3%). These categories were more likely than others to feature articles published in journals with rnIF .901 or higher. In some cases, it might even be more insightful to consider the distribution of both the .801 to .900 and the .901 and higher together. For example, 88.2% of all Space Science articles were published in journals with rnIF of .801 or higher. In most cases, but not all, the

Table 12. Cross Tabulation for rnIF

Category		**.600 or less**	**.601 to .800**	**.801 to .900**	**.901 or higher**	**Total**
Materials Science	Count	225	703	569	1893	3390
	% within Category	6.6%	20.7%	16.8%	55.8%	100.0%
	% within All	2.1%	5.3%	3.2%	8.4%	5.3%
	Adjusted Residual	**-15.8**	-.1	**-14.3**	**25.7**	
Multidisciplinary	Count	2	230	779	1434	2445
	% within Category	.1%	9.4%	31.9%	58.7%	100.0%
	% within All	.0%	1.7%	4.4%	6.3%	3.8%
	Adjusted Residual	**-22.3**	**-14.2**	**5.0**	**24.7**	
Chemistry	Count	1012	1592	2629	3832	9065
	% within Category	11.2%	17.6%	29.0%	42.3%	100.0%
	% within All	9.6%	11.9%	14.9%	16.9%	14.1%
	Adjusted Residual	**-14.6**	**-8.2**	**3.5**	**15.0**	
Plant & Animal Science	Count	750	555	672	1665	3642
	% within Category	20.6%	15.2%	18.5%	45.7%	100.0%
	% within All	7.1%	4.2%	3.8%	7.4%	5.7%
	Adjusted Residual	**7.0**	**-8.5**	**-12.6**	**13.6**	
Geosciences	Count	398	899	1028	1853	4178
	% within Category	9.5%	21.5%	24.6%	44.4%	100.0%
	% within All	3.8%	6.7%	5.8%	8.2%	6.5%
	Adjusted Residual	**-12.5**	1.2	**-4.3**	**12.7**	
Space Science	Count	82	467	2093	2028	4670
	% within Category	1.8%	10.0%	44.8%	43.4%	100.0%
	% within All	.8%	3.5%	11.9%	9.0%	7.3%
	Adjusted Residual	**-28.1**	**-18.9**	**27.6**	**12.1**	

Table 12 (Continued)

Category		**.600 or less**	**.601 to .800**	**.801 to .900**	**.901 or higher**	**Total**
Social Sciences, General	Count	60	47	96	301	504
	% within Category	11.9%	9.3%	19.0%	59.7%	100.0%
	% within All	.6%	.4%	.5%	1.3%	.8%
	Adjusted Residual	-2.8	**-6.4**	**-4.3**	**11.5**	
Environment / Ecology	Count	545	522	756	1284	3107
	% within Category	17.5%	16.8%	24.3%	41.3%	100.0%
	% within All	5.2%	3.9%	4.3%	5.7%	4.8%
	Adjusted Residual	1.7	**-5.6**	**-4.0**	**7.2**	
Agricultural Sciences	Count	22	48	39	92	201
	% within Category	10.9%	23.9%	19.4%	45.8%	100.0%
	% within All	.2%	.4%	.2%	.4%	.3%
	Adjusted Residual	-2.1	1.1	-2.6	**3.1**	
Engineering	Count	948	934	890	1492	4264
	% within Category	22.2%	21.9%	20.9%	35.0%	100.0%
	% within All	9.0%	7.0%	5.0%	6.6%	6.6%
	Adjusted Residual	**10.6**	1.8	**-10.0**	-.4	
Psychiatry / Psychology	Count	29	25	43	41	138
	% within Category	21.0%	18.1%	31.2%	29.7%	100.0%
	% within All	.3%	.2%	.2%	.2%	.2%
	Adjusted Residual	1.5	-.8	1.0	-1.4	
Pharmacology & Toxicology	Count	44	65	39	56	204
	% within Category	21.6%	31.9%	19.1%	27.5%	100.0%
	% within All	.4%	.5%	.2%	.2%	.3%
	Adjusted Residual	2.0	**3.9**	-2.7	-2.3	

Table 12 (Continued)

Category		**.600 or less**	**.601 to .800**	**.801 to .900**	**.901 or higher**	**Total**
Neuroscience & Behavior	Count	185	150	84	177	596
	% within Category	31.0%	25.2%	14.1%	29.7%	100.0%
	% within All	1.8%	1.1%	.5%	.8%	.9%
	Adjusted Residual	**9.7**	2.6	**-7.4**	-2.9	
Immunology	Count	19	28	10	2	59
	% within Category	32.2%	47.5%	16.9%	3.4%	100.0%
	% within All	.2%	.2%	.1%	.0%	.1%
	Adjusted Residual	**3.3**	**5.0**	-1.8	**-5.1**	
Clinical Medicine	Count	61	128	99	82	370
	% within Category	16.5%	34.6%	26.8%	22.2%	100.0%
	% within All	.6%	1.0%	.6%	.4%	.6%
	Adjusted Residual	.0	**6.6**	-.3	**-5.3**	
Economics & Business	Count	64	70	34	33	201
	% within Category	31.8%	34.8%	16.9%	16.4%	100.0%
	% within All	.6%	.5%	.2%	.1%	.3%
	Adjusted Residual	**5.9**	**4.9**	**-3.4**	**-5.6**	
Microbiology	Count	189	181	55	121	546
	% within Category	34.6%	33.2%	10.1%	22.2%	100.0%
	% within All	1.8%	1.4%	.3%	.5%	.9%
	Adjusted Residual	**11.5**	**7.1**	**-9.1**	**-6.4**	
Molecular Biology & Genetics	Count	346	490	218	177	1231
	% within Category	28.1%	39.8%	17.7%	14.4%	100.0%
	% within All	3.3%	3.7%	1.2%	.8%	1.9%
	Adjusted Residual	**11.2**	**16.6**	**-7.8**	**-15.5**	

Table 12 (Continued)

Category		**.600 or less**	**.601 to .800**	**.801 to .900**	**.901 or higher**	**Total**
Computer Science	Count	1365	796	1016	996	4173
	% within Category	32.7%	19.1%	24.3%	23.9%	100.0%
	% within All	12.9%	6.0%	5.8%	4.4%	6.5%
	Adjusted Residual	**29.3**	-2.8	**-4.7**	**-16.0**	
Physics	Count	1181	2902	4553	3153	11789
	% within Category	10.0%	24.6%	38.6%	26.7%	100.0%
	% within All	11.2%	21.7%	25.8%	13.9%	18.4%
	Adjusted Residual	**-20.8**	**11.3**	**30.0**	**-21.5**	
Biology & Biochemistry	Count	652	903	485	339	2379
	% within Category	27.4%	38.0%	20.4%	14.2%	100.0%
	% within All	6.2%	6.8%	2.8%	1.5%	3.7%
	Adjusted Residual	**14.7**	**21.0**	**-7.9**	**-21.9**	
Mathematics	Count	2371	1617	1447	1591	7026
	% within Category	33.7%	23.0%	20.6%	22.6%	100.0%
	% within All	22.5%	12.1%	8.2%	7.0%	10.9%
	Adjusted Residual	**41.5**	4.8	**-13.7**	**-23.5**	
Total	Count	10550	13352	17634	22642	64178
	% within Category	16.4%	20.8%	27.5%	35.3%	100.0%
	% within All	100.0%	100.0%	100.0%	100.0%	100.0%

categories with high distributions in the top rnIF range were also less likely to publish in the lowest range of rnIF. For example, only 6.6% of observations from within Material Science and only 1.8% of Space Science articles fell into the .600 or lower rnIF range, versus the overall of 16.4%.

By contrast, some categories were more likely than others to contain articles published in lower ranking journals, respective to their domains. Only 22.6% of articles in Mathematics journals were published in journals with rnIF .901 or higher, compared to the 35.3% overall. Furthermore, 33.7% of Mathematics articles were published in journals with rnIF .600 or lower, compared to the 16.4% overall. The adjusted residual value for Mathematics in the .600 or less range is the most extreme of all residuals in the cross-tabulation, giving us a sense of just how meaningful this disproportionately high number of publications in low-impact journals is. Likewise, Computer Science articles were less likely, compared to others, to have been published in the highest impact journals for this field (only 23.9% to the overall 35.3%), while much more likely to publish in the lowest impact journals (32.7% versus the 16.5% overall). Attention is also called to Biology & Biochemistry, for which a total of 65.4% of articles were published in journals with rnIF less than .801, and only 34.6% of articles were published in the two higher ranges of journals with rnIF .801 or greater.

Discussion

The results demonstrated that, by linking awards to publications, it is possible to learn something about the time it takes to publish research after it has been funded. Acknowledgement lag peaked at 4 years and experienced an increasingly sharp drop-off after that point. One of the unexpected findings in these results was that some articles continued to acknowledge grants as long as 30 years or more past. Although the plotted frequency distribution shows these are very, very few in number (only 25 out of 64,178 had a lag of 30 or more) it is interesting that they exist at all, and it would be a prospect for future research to learn more about why researchers may or may not choose to acknowledge an award even after the official grant period has ended. Five articles with 30+ years acknowledgement lag in in the Chemistry category referenced NSF award number 7904825, which was titled “Purchase of a High-Field Multinuclear Nuclear Magnetic Resonance Spectrometer”, while five others in Chemistry with equally as long lag referenced award number 8018643, which was titled, “Ft-Nmr Instrumentation for Research”. Keeping in mind that these 10 cases are very unusual for Chemistry as a research category, given that Chemistry was found to most often publish with acknowledgements within 2 years or less, in these cases the researchers would appear to be continuing to acknowledge the grants that provided scientific instruments still in use. Another case of a long acknowledgment lag time was an article titled “Rebuilding after collapse: evidence for long-term cohort dynamics in the native Hawaiian rain forest”, which acknowledged award number 7910993 from 1979. The abstract of the article describes analysis of forest canopy over a period of 27 years. It is not clear from the abstract the role that the grant played, which could have been that it contributed to the development of the model employed, the establishment of the original forest plots in the seventies and eighties, or some other purpose.

The results also demonstrated clear differences between research fields in terms of acknowledgement lag. Data for some sciences such as Physics, Chemistry, and Space Science, in addition to Mathematics tended to have shorter award-to-publication lag times compared to other categories such as Plant & Animal Sciences or Social Sciences. Although it is impossible to know for certain without further investigation, we can consider some factors which might possibly contribute to these differences. One factor could be differences in the publication lag from one category of research to the next. As previously noted, publication lag is a measure of the difference between the point in time in which a manuscript was submitted to the point in time which it is published. Previous studies, such as one by Björk and Solomon (2013), have found differences to exist in the typical publication lag for various categories of research. It may not be a perfect explanation because all of their results do not translate exactly to the categories used or findings in the current study, but there are a few similarities. For example, they do find that Chemistry and Physics tend to have a shorter publication lag than Social Science or Economics, which is consistent with the results shown for acknowledgement lag in this research.

Another potential contributing factor may be the availability of "Letters" journals, a category of publications which, "has come into existence exclusively for the rapid publication of preliminary results of research," and which have "succeeded in appreciably speeding up dissemination of results of research" (Subramanyam, 1981). If certain fields of research have more venues to participate in rapid communication of findings, in addition to a culture which expects as much, then this may serve as a contributing factor to shorter acknowledgement times.

A final possibility to consider may simply be the nature of the research being performed. Longitudinal studies of people over time in Social Sciences, or studies of how plants and living things change over time, might in some cases take longer to publish findings than, for example, a molecular simulation or neutron scattering experiment where the data is immediately available for analysis.

In addition to the results about acknowledgement lag, the results regarding publication impact also showed meaningful differences in how individual categories contributed to the overall picture of research acknowledging NSF support. The fact that so many articles published in Mathematics or Computer Science journals fall within the lower ranking rnIF is not negligible, considering that these observations constituted 10.9% and 6.5% of the overall dataset, respectively. However, it is perhaps more difficult to speculate as to the contributing factors for these differences, lacking internal knowledge of the NSF's program management. One of the more readily available sources of information we have about how the NSF prioritizes research and development is the record of research and development (R&D) expenditures by agency, via the SEI reports, which use an 8 category schema: Life sciences, Engineering, Physical sciences, Environmental sciences, Math/Computer sciences, Psychology, Social sciences, and Other sciences. Computer science/Math was the largest research expenditure area in 2013 and the second largest in 2011 (National Science Board, 2014, 2016). Yet the publication data for the current dataset shows these to be some of the least impressive research areas according to the measures employed in this study. As previously noted, journal impact factor is not a measure without controversy, and so perhaps some programs prioritize journal impact factor as a measure of performance more so than others. What this research provides, however, is a systematic and consistent method for comparing the publication outcomes of different research areas, which has the potential to be useful from an administrative perspective.

It was also interesting to observe that one of the highest performing categories was Multidisciplinary. Although Multidisciplinary as a category does not really tell us anything about differentiating between domains, this group includes very noteworthy journals such as *Nature* and *Proceedings of the National Academy of Sciences.* The complications presented by multidisciplinary journals for this kind of analysis are further discussed in the Limitations section.

A final point of discussion is the fact that over 59% of all Social Sciences publications were published in journals with rnIF .901 or higher. This was a small category, constituting less than 1% of the total dataset, but the literature had made it clear that high retrieval rates should not be expected for Social Sciences. Also, Social Science is not typically a high spending priority for the NSF, with only Psychology receiving less funding according to the 2014 and 2016 SEI. This study was exploratory, and no hypotheses were offered as to which categories, if any, would have the highest proportion of high-impact journal publications. Nonetheless, it is interesting to notice that research priorities in terms of overall spending do not necessarily indicate which areas of research will produce the highest quality of outcomes.

Chapter 5

Conclusions and Recommendations

In conclusion, this research has demonstrated that funding agencies have much to gain from linking grant awards as inputs to scholarly publications as outputs. Through systematic analysis of data from one source which has been enriched by the other, the outputs were able to be analyzed in relation to the inputs. The overall acknowledgement lag has been described for research articles supported by the National Science Foundation, and important domain-differences in this measure have mean recognized. Some research areas, such as Plant & Animal Science or Social Science, were found to be more likely to have an acknowledgement lag of 7 years or more, whereas other research areas, such as Physics and Chemistry, were most likely to publish with an acknowledgement in 2 years or less. Data-driven understanding of differences in expected norms for various categories of research is a potentially valuable source of information.

This research has also identified differences in journal impact factor for publications supported by the NSF. Materials Science was found to be one of the strongest areas of research impact, tending to publish in high-quality journals. Social Science articles were far fewer in number, but almost equally likely to be published in top journals. Although Physics research supported by the NSF was less likely than other categories to be published in the very highest ranking journals, it still performed well in the second highest tier of impact. It was also found that when NSF supported researchers did publish in Multidisciplinary journals, the clear majority of these articles tended to be published in elite multidisciplinary journals. Computer Science and Mathematics were found to be more likely than other categories to publish in lower-impact journals rather than high-impact journals. Overall, this research has demonstrated a data-processing strategy based on rank-normalized impact factor for systematically comparing the quality of research produced across categories.

In many ways, the ability to conduct such research is becoming possible thanks to the increasing willingness of publishers and databases to recognize funding acknowledgements as a value-added and meaningful enrichment to bibliographic data. More consistent efforts to identify and document funding acknowledgements will only serve to enhance the possibilities and effectiveness of input-output studies which link sources of funding to publication outcomes.

Implications

This project developed a methodology for assessment of research which acknowledges support by a funding agency, requiring minimal manual intervention. Implications for methodology include that the workflow serves as a kind of proof of concept for the development of a fully automated analytical system. Implemented as a live analytical processing system, the use of a bibliographic database’s API service could be scripted to perform weekly or monthly checks for new data meeting the information retrieval criteria. As new Journal Citation Reports are made available on an annual basis, these too can be loaded. The result would be a dynamic analytical system contributing to the NSF or any other agency’s ability to track outcomes and trends related to their investments in a way that does not depend solely on self-reported data from investigators, and in a way which leverages the high-quality of data curation provided by bibliographic databases. Changes to distribution of impact factor of publications is something that could be tracked over time. The design of this workflow is furthermore modular enough that other kinds of metrics besides impact factor could be plugged into the design as well.

One of the more interesting implications of the findings is tied to the discovery that awards from so far back continued to be acknowledged in recent papers. Although these were few in number, the fact that it was found at all came as something of a surprise. Particularly in cases where the award contributed to the development or acquisition of scientific instruments or other resources, it demonstrates that these funds continued to generate returns on investment far after the award was given. For an agency which is often in a position of having to justify why they should be budgeted money for future research, it is advantageous to be able to show this kind of proof of the long-term value created.

The significance of the findings as to acknowledgement lag are also useful in terms of showing return on investment to both internal and external stakeholders. With this information, agencies can have a better understanding how long it reasonably takes for results to be published. The difference between having that information exist as tacit knowledge scattered across various program administrators, directorate managers, and subject matter experts, versus the systematic collection and consistent analysis of data across the entire organization, is that the latter is able to be succinctly summarized and used. It is actionable information at the strategic level.

Lessons Learned

In any situation where a formal solution were to be developed from a prototype, the database design, queries, and code would be thoroughly reviewed for efficiency and optimization. But there are some points beyond such generalities at which the process could be improved upon. One of the most obvious examples is that slightly more than 4% of the Journal Citation Report data rows failed to map to a category using the ESI mapping table. 3.8 percent of article-to-award relationships were consequently discarded, being excluded from the analysis due to their inability to be classified within the schema. This study did not examine the causes for why a journal failed to map to the category schema. However, the situation highlights the importance of consistency across data products for development of this type of analytical system. One possible contributing factor to mapping failures may have been the fact that title changes are common for serial publications. The ESI schema lists a category mapping for the title version most current in relation to the release date of the mapping table version, meaning that past title versions would not be categorized. For example, the journal *Surgical Neurology* became *World Neurosurgery* in 2010. The most current ESI category table has a mapping entry for *World Neurosurgery*, but not *Surgical Neurology*. Any articles published under the previous source title version would not be able to be classified. To solve this type of problem, bibliographic databases should provide category mapping data for all past and present name variants of journals, thus avoiding an unnecessary loss of data. The need for mapping of all name variants will hold true for any classification schema, not just the ESI schema.

A second opportunity for improvement would rely on the cooperation of the vendor for Journal Citation Reports to provide this data in a cleaner, more raw format. The JCR data in this case did require some manual intervention. However, this sort of data cleaning and preparation is really something that should be handled on the vendor's side, and is more a matter of formatting than anything to prepare it for ingestion into a MySQL database. The solution for incorporating JCR data as an element of an automated analytical system need not be anything more complicated than a simple scheduled deposit on an FTP server on the vendor's part, and a server job scanning for new deposits on the agency's part.

Limitations

One of the challenges encountered in interpreting categorical differences in lag times was the very large differences in category sizes. Some categories contained thousands of observations while others contained less than one or two-hundred. This could have occurred for several reasons. The categorization outcomes could reflect real differences in the investment of or management by the NSF regarding different research domains. It is also possible that number of observations in a given category could be influenced by differences in how often researchers in these areas tend to publish, or when and how researchers acknowledge grants. Most likely, the categorization schema itself also influences the size of categories. When and how one chooses to divide a larger category into smaller categories will naturally affect the size of those categories. For example, Agricultural Sciences only had 201 lag observations, but had the schema combined these journals with those in Plant & Animal Sciences, they would have been part of a larger size of category. The data for a category such as Immunology was very small (58 observations) compared to others categories which had thousands of observations.

Additionally, the categorization occurred at the journal level rather than the article level, which in most cases still produced interesting findings. However, some information was lost by doing so, especially when we consider the fact that one of the highest performing “categories” in terms of articles published in the most elite journals was the Multidisciplinary group. This problem of journal-level classification meant that 2,445 article, or 3.8% of the dataset, could not be associated with the appropriate area of research, which is a meaningful limitation.

The decision to use an existing, journal-based category schema for this project was made for reasons of compatibility with the bibliographic data being analyzed, for purposes of consistency, and for simple convenience. However, the role of the categorization schema in influencing outcomes merits further review.

An additional limitation of the study is the likelihood of incomplete data. As earlier noted, a very recent study Tang et al. (2016) found that for bibliographic data of publications published between 2009 and 2014, only 47% of publication records in the SCIE and 16% of SSCI records contain a funding acknowledgement. Based on this, we may expect that Social Science data specifically was underrepresented, and that funding acknowledgements may not have been consistently identified by Web of Science. However, it is possible that for some of these, there was no relationship to a funding agency to report.

Future Directions

Although the motivations for and context in which a grant is acknowledged was outside the scope of this work, the wide range of acknowledgement lag times, in addition to the differences in lag tendencies amongst categories, makes the case for seeking a better understanding of how and when researchers make acknowledgements to funding agencies. It should also be investigated how resources originating from a grant may continue to be employed long after the grant period has ended. As previously noted, agencies have a vested interest in recognizing where and what kinds of awards continue to generate long-term returns on investments. Better understanding of these factors will contribute to improved methodology for performing systematic, large-scale analysis and reporting on federally funded research outcomes.

An additional possibility for extending this research would be to investigate if there is a correlation between the amount of an award and the impact or number of publications produced which acknowledge the grant. There are a few reasons why the dataset in this study was not suitable for answering such a research question. First, it only included a five-year span of publications between 2010 and 2014 (stopping at 2014 due to that this was the most recent availability of Journal Citation Reports at the time of the analysis). In fact, we could expect to start finding complete years' worth of funding acknowledgements in the Web of Science no sooner than 2009 at the earliest. What we have learned from the current study is that some domains are more likely to publish after longer periods of time than others, after receiving a grant. An award granted to researchers in Plant in Animal Sciences in 2000 would have had 10 or more years to show progress in terms of publication outputs, but the funding acknowledgement would not have been included in the bibliographic data any sooner than late 2008 or 2009 at the earliest. An award granted to researchers who publish in Plant & Animal Sciences in 2010 would not have had reasonable enough time to produce publications, based on an expectation that they are more likely to take seven or more years to publish than other categories. As the number of years of bibliographic data which include funding acknowledgements grows, such analysis would be less likely to be biased in favor of some domains versus others.

The second problem with using this dataset to test for a correlation between funding amount and output, particularly across categories, is a methodological one. The categorization schema used for mapping the grant-article relationships was based on the journal of the article's publication. To compute aggregate statistics for a grant and compare these across categories would require that the categorization be applied to the grant rather than the article. This is because we cannot guarantee that multiple articles from the same grant will always publish within the same category of journal. As an answer to this problem, a future study might consider using the division within the granting agency as an indication of research category (e.g., Division of Physics or Division of Molecular and Cellular Bioscience). The grant-article relationships evaluated in the current research were spread out over awards from 40 different NSF divisions—a large enough number of categories to present its own analytical challenges, but an alternative none the less.

An additional promising direction for further investigation would be to conduct the same analysis for different funding agencies. We might look to see if other funding agencies, with their varying missions, manifest different outcomes in terms of strengths and weaknesses in publication impact. It is also necessary to do additional studies with new data in order for the acknowledgement lag patterns to be validated. Analyzing other agencies would both fulfill this need and would also serve to confirm whether or not these patterns are universal to their categories or research, or if there is something within the operations of the NSF itself that influences these outcomes.

List of References

Althouse, B. M., West, J. D., Bergstrom, C. T., & Bergstrom, T. (2009). Differences in Impact Factor Across Fields and Over Time. *Journal of the American Society for Information Science and Technology, 60*(1), 27-34. doi:10.1002/asi.20936

Auranen, O., & Nieminen, M. (2010). University research funding and publication performance—An international comparison. *Research Policy*, *39*(6), 822-834. doi:http://dx.doi.org/10.1016/j.respol.2010.03.003

Belter, C. W. (2013). A bibliometric analysis of NOAA's Office of Ocean Exploration and Research. *Scientometrics, 95*(2), 629-644. doi:10.1007/s11192-012-0836-0

Björk, B.-C., & Solomon, D. (2013). The publishing delay in scholarly peer-reviewed journals. *Journal of Informetrics*, *7*(4), 914-923.

Bordons, M., Fernández, M., & Gómez, I. (2002). Advantages and limitations in the use of impact factor measures for the assessment of research performance. *Scientometrics*, *53*(2), 195-206.

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45-80.

Boyack, K. W., & Jordan, P. (2011). Metrics associated with NIH funding: a high-level view. *Journal of the American Medical Informatics Association*, *18*(4), 423-431. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3128410/pdf/amiajnl-2011-000213.pdf

Broadus, R. (1987). Toward a definition of "bibliometrics". *Scientometrics, 12*(5-6), 373-379.

Coppin, A. (2013). Finding Science and Engineering Specific Data Set Usage or Funding Acknowledgements. *Issues in Science and Technology Librarianship*.

De Bellis, N. (2009). *Bibliometrics and citation analysis: from the science citation index to cybermetrics*: Scarecrow Press.

Diodato, V. (1994). *Dictionary of Bibliometrics*. Binghamton, NY: Haworth Press.

Garfield, E. (1999). Journal impact factor: a brief review. *Canadian Medical Association Journal, 161*(8), 979-980. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1230709/pdf/cmaj_161_8_979.pdf

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H., & Terliesner, J. (2014). Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics, 101*(2), 1145-1163.

Hoeffel, C. (1998). Journal impact factors. *Allergy, 53*(12), 1225-1225.

Jongbloed, B., & Vossensteyn, H. (2001). Keeping up Performances: an international survey of performance-based funding in higher education. *Journal of Higher Education Policy & Management, 23*(2), 127-145. doi:10.1080/13600800120088625

Jowkar, A., Didegah, F., & Gazni, A. (2011, 2011). *The effect of funding on academic research impact: a case study of Iranian publications.* Paper presented at the Aslib Proceedings.

Leydesdorff, L. (2012). Alternatives to the journal impact factor: I3 and the top-10%(or top-25%?) of the most-highly cited papers. *Scientometrics, 92*(2), 355-365.

Leydesdorff, L., & Opthof, T. (2010). Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. *Journal of the American Society for Information Science and Technology, 61*(11), 2365-2369.

MacLean, M., Davies, C., Lewison, G., & Anderson, J. (1998). Evaluating the research activity and impact of funding agencies. *Research Evaluation, 7*(1), 7-16. Retrieved from http://rev.oxfordjournals.org/content/7/1/7.abstract

Mongeon, P., Brodeur, C., Beaudry, C., & Larivière, V. (2015). *On Decreasing Returns to Scale in Research Funding.* Paper presented at the 15th International Conference on Scientometrics & Infometrics, Istanbul, Turkey.

National Science Board. (2014). *Science and Engineering Indicators*. (NSB 14-01). Arlington VA: National Science Foundation.

National Science Board. (2016). *Science and Engineering Indicators 2016*. Retrieved from Arlington, VA: http://www.nsf.gov/statistics/2016/nsb20161/#/downloads/report

National Science Foundation. (n.d.). About the National Science Foundation.   Retrieved from http://www.nsf.gov/about/

National Science Foundation Act of 1950, Pub. L. No. 507 § 247 (1950).

Piwowar, H. (2013). Altmetrics: Value all research products. *Nature, 493*(7431), 159-159.

Pritchard, A. (1969). Documentation Notes. *Journal of Documentation, 25*(4), 344-349. doi:10.1108/eb026482

Pudovkin, A. I., & Garfield, E. (2004). Rank-normalized impact factor: A way to compare journal performance across subject categories. *Proceedings of the American Society for Information Science and Technology, 41*(1), 507-515. doi:10.1002/meet.1450410159

Racki, G. (2009). Rank-normalized journal impact factor as a predictive tool. *Archivum immunologiae et therapiae experimentalis*, *57*(1), 39-43.

Rigby, J. (2011). Systematic grant and funding body acknowledgement data for publications: New dimensions and new controversies for research policy and evaluation. *Research Evaluation, 20*(5), 365-375.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ : British Medical Journal, 314*(7079), 498-502. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2126010/pdf/9056804.pdf

Subramanyam, K. (1981). *Scientific and Technical Information Resources*. New York, NY: Marcel Dekker, Inc.

Tang, L., Hu, G., & Liu, W. (2016). Funding acknowledgment analysis: Queries and Caveats. *arXiv preprint arXiv:1601.00245*.

Thomson Reuters. (2009). Funding Acknowledgements. Retrieved from http://wokinfo.com/products_tools/multidisciplinary/webofscience/fundingsearch/

Thomson Reuters. (2015). *Research Area Schemas*. Retrieved from http://ipscience-help.thomsonreuters.com/inCites2Live/researchAreaSchema.pdf

Verma, I. M. (2015). Impact, not impact factor. *Proceedings of the National Academy of Sciences*, *112*(26), 7875-7876. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4491750/pdf/pnas.201509912.pdf

Wang, J., & Shapira, P. (2015). Is there a relationship between research sponsorship and publication impact? An analysis of funding acknowledgments in nanotechnology papers. *PloS one, 10*(2), e0117727.

Zhao, D. Z. (2010). Characteristics and impact of grant-funded research: a case study of the library and information science field. *Scientometrics, 84*(2), 293-306. doi:10.1007/s11192-010-0191-y

Appendices

Appendix A

Code for Categories & Computing rnIF

SQL to code journals in a JCR table with categories:

```
UPDATE JCR2014 JOIN ESI2016
ON JCR2014.FullJournalTitle = ESI2016.FullTitle
SET JCR2014.CategoryName = ESI2016.CategoryName;
```

SQL to summarize outcomes of mapping process:

```
SELECT CategoryName, Count(CategoryName)
FROM JCR2014
WHERE CategoryName IS NOT NULL
GROUP BY CategoryName
;
```

SQL to count number of JCR Journal entries in a year that failed to map to a category:

```
SELECT COUNT(*) FROM JCR2014 WHERE CategoryName IS NULL;
```

SQL to select categorized journal data from JCR table, ordered by category and then impact factor descending:
(to be exported as JSON)

```
Select FullJournalTitle, JournalIF, JCRYear,CategoryName,
RankInCategory, rnIF
FROM JCR2014
WHERE CategoryName IS NOT NULL
ORDER BY CategoryName, JournalIF DESC;
```

Excerpt of JSON export:

(Note that unpopulated columns CategoryName, RankInCategory, and rnIF had been added to the JCR data table prior to export. Also, a quirk of MySQL Workbench (v 6.3.6 build 511 CE) is that empty fields are populated as NULL in the JSON file, which is not actually valid JSON. The JSON file must be opened in a text editor and all instances of NULL replaced with a valid form such as "" or "NULL" before it can be decoded.)

```
	{
		"FullJournalTitle" : "PHLEBOLOGIE-ANNALES VASCULAIRES",
		"JournalIF" : 0,
		"JCRYear" : 2012,
		"CategoryName" : "NULL",
		"RankInCategory" : "NULL",
		"rnIF" : "NULL"
	},
	{
		"FullJournalTitle" : "TEMPO PSICANALITICO",
		"JournalIF" : 0,
		"JCRYear" : 2012,
		"CategoryName" : "NULL",
		"RankInCategory" : "NULL",
		"rnIF" : "NULL"
	},
	{
		"FullJournalTitle" : "ANNUAL REVIEW OF NUTRITION",
		"JournalIF" : 9.158,
		"JCRYear" : 2012,
		"CategoryName" : "AGRICULTURAL SCIENCES",
		"RankInCategory" : "NULL",
		"rnIF" : "NULL"
	},
	{
		"FullJournalTitle" : "NUTRITION RESEARCH REVIEWS",
		"JournalIF" : 5.5,
		"JCRYear" : 2012,
		"CategoryName" : "AGRICULTURAL SCIENCES",
		"RankInCategory" : "NULL",
		"rnIF" : "NULL"
	},
```

File Contents for rnIF.php:

The following script expects a json export of a specific year's JCR data as described in the text. The script could easily be modified to accept another format of data so long as the data were still loaded to an array. The script loads the file and determines the number of journals in each category for that year's JCR data. It will then assign each journal its rank within its respective category, and compute the rank-normalized impact factor based on that rank. Finally, that journal's record will be updated in the database with the newly computed values.

Note: The ***filename*** *for the data and the database* ***table name*** *must be updated for each year processed, because it assumes that each year's JCR data had been loaded to its own table.*

```
<?php
      $errorFile = 'errors.txt';
      $servername = "host:port";
      $username = "username";
      $password = "password";
      $dbname = "database";
      // Create connection
      $connection = new mysqli($servername, $username, $password,
$dbname);
      // Check connection
      if ($connection->connect_error) {
            die("Connection failed: " . $connection->connect_error);
      }
            echo("\nDatabase opened.\n");

//UPDATE THIS FILE NAME FOR THE DATA BEING PROCESSED
$string = file_get_contents("JCR2014.json");
$JCRarray = json_decode($string, true);
$length=count($JCRarray);
echo("Total Journal Records: $length \n");
$trackCategories=array();

$i=0;
while ($i<$length) {
      $setSize = 1;
      $ii=$i; //inner loop management
      while
($JCRarray[$ii]['CategoryName']==$JCRarray[$ii+1]['CategoryName']){
            $JCRarray[$ii]['RankInCategory']=$setSize;
            $setSize = $setSize+1;
            $ii=$ii+1;
      }
      if
($JCRarray[$ii]['CategoryName']!=$JCRarray[$ii+1]['CategoryName']){
            $JCRarray[$ii]['RankInCategory']=$setSize;

            $setStartPos = $ii-($setSize-1);
            $setCurrentPos = $setStartPos;
            $setEndPos = $setStartPos+($setSize-1);
            while ($setCurrentPos<=$setEndPos){
                  $JCRarray[$setCurrentPos]['rnIF']=(($setSize-
$JCRarray[$setCurrentPos]['RankInCategory']+1)/$setSize);
                  $RankInCategory =
$JCRarray[$setCurrentPos]['RankInCategory'];
                  $rnIF = $JCRarray[$setCurrentPos]['rnIF'];
                  $FullJournalTitle =
$JCRarray[$setCurrentPos]['FullJournalTitle'];


//UPDATE THIS TABLE NAME FOR THE DATA BEING PROCESSED

$sql ="UPDATE JCR2014 SET RankInCategory = $RankInCategory, rnIF=$rnIF
            WHERE FullJournalTitle='$FullJournalTitle'";

      if ($connection->query($sql) === TRUE) {
      echo "Record updated.\n";
      }
      else {
      echo "Error: " . $sql . "\n" . $connection->error;
      file_put_contents($errorFile, "Error: " . $sql . "\n" .
$connection->error, FILE_APPEND);}

            $setCurrentPos++;
      }

      }
      $category=$JCRarray[$i]['CategoryName'];
      $trackCategories[]="$category: $setSize";
      $i=$i+$setSize;
}
print_r($trackCategories);
$connection->close();
echo("\nDatabase closed.\n");

?>
```

Example Categorization Output of rnIF.php:

The reported count for each category within a given year's JCR data is output by the rnIF.php script, confirming what is seen in the database with an sql query.

```
Total Journal Records: 11161
Array
(
    [0] => AGRICULTURAL SCIENCES: 320
    [1] => BIOLOGY & BIOCHEMISTRY: 404
    [2] => CHEMISTRY: 508
    [3] => CLINICAL MEDICINE: 1815
    [4] => COMPUTER SCIENCE: 365
    [5] => ECONOMICS & BUSINESS: 532
    [6] => ENGINEERING: 812
    [7] => ENVIRONMENT/ECOLOGY: 318
    [8] => GEOSCIENCES: 385
    [9] => IMMUNOLOGY: 153
    [10] => MATERIALS SCIENCE: 332
    [11] => MATHEMATICS: 480
    [12] => MICROBIOLOGY: 112
    [13] => MOLECULAR BIOLOGY & GENETICS: 289
    [14] => Multidisciplinary: 39
    [15] => NEUROSCIENCE & BEHAVIOR: 326
    [16] => PHARMACOLOGY & TOXICOLOGY: 261
    [17] => PHYSICS: 296
    [18] => PLANT & ANIMAL SCIENCE: 744
    [19] => PSYCHIATRY/PSYCHOLOGY: 602
    [20] => SOCIAL SCIENCES, GENERAL: 1871
    [21] => SPACE SCIENCE: 53
)
```

Appendix B

Code for Award & Bibliographic Data

An example of an NSF award record is as follows. Please note that the abstract has been intentionally shortened in this example for space considerations:

1100080.xml (Example NSF Award Record)

```
<?xml version="1.0" encoding="UTF-8"?>

<rootTag>
  <Award>
    <AwardTitle>Surface Science and Engineering Towards
Bioactive Bulk Metallic Glasses</AwardTitle>
    <AwardEffectiveDate>06/01/2011</AwardEffectiveDate>
    <AwardExpirationDate>05/31/2016</AwardExpirationDate>
    <AwardAmount>296536</AwardAmount>
    <AwardInstrument>
      <Value>Standard Grant</Value>
    </AwardInstrument>
    <Organization>
      <Code>07030000</Code>
      <Directorate>
        <LongName>Directorate For Engineering</LongName>
      </Directorate>
      <Division>
        <LongName>Div Of Civil, Mechanical, &amp; Manufact
Inn</LongName>
      </Division>
    </Organization>
    <ProgramOfficer>
      <SignBlockName>Alexis Lewis</SignBlockName>
    </ProgramOfficer>
    <AbstractNarration>The research objective of this
award is to elucidate the mechanisms underlying the ion
beam interactions with the bulk metallic glasses (BMGs)
and the impacts of ion implantation on the surface
bioactivity of BMGs. The research will (i) identify the
key processes and variables for ion implantation towards
bioactive BMGs, (ii) investigate the effects of ion
implantation on both surface and electrochemical
```

```
properties of BMGs, and (iii) study the biological
activity of ion implanted BMGs, with….</AbstractNarration>
    <MinAmdLetterDate>05/13/2011</MinAmdLetterDate>
    <MaxAmdLetterDate>05/13/2011</MaxAmdLetterDate>
    <ARRAAmount/>
    <AwardID>1100080</AwardID>
    <Investigator>
      <FirstName>Peter</FirstName>
      <LastName>Liaw</LastName>
      <EmailAddress>pliaw@utk.edu</EmailAddress>
      <StartDate>05/13/2011</StartDate>
      <EndDate/>
      <RoleCode>Co-Principal Investigator</RoleCode>
    </Investigator>
    <Investigator>
      <FirstName>Wei (Lydia)</FirstName>
      <LastName>He</LastName>
      <EmailAddress>whe5@utk.edu</EmailAddress>
      <StartDate>05/13/2011</StartDate>
      <EndDate/>
      <RoleCode>Principal Investigator</RoleCode>
    </Investigator>
    <Institution>
      <Name>University of Tennessee Knoxville</Name>
      <CityName>KNOXVILLE</CityName>
      <ZipCode>379960003</ZipCode>
      <PhoneNumber>8659743466</PhoneNumber>
      <StreetAddress>1 CIRCLE PARK</StreetAddress>
      <CountryName>United States</CountryName>
      <StateName>Tennessee</StateName>
      <StateCode>TN</StateCode>
    </Institution>
  </Award>
</rootTag>
```

The processing code is organized into four files, which should be in the same directory:

- ***main.php*** - Should be executed as the starting point of the program.
- ***database_connection.php*** **-** Handles opening and closing the connection to the database.
- ***process_NSF_awards.php*** – Parses the XML data of the awards file and handles construction of any queries for inserting awards data.
- ***WoS_manage_sessions.php*** - which handles creating and closing the sessions for communicating with the Web of Science API.
- ***WoS_execute_search.php*** - controls actual searching, retrieval, and processing of bibliographic records extracted through the Web of Science API.

File Contents for main.php

```
<?php
$time_start = microtime(true);
include('database_connection.php');
include('WoS_manage_sessions.php');
include('process_NSF_awards.php');
include('WoS_execute_search.php');
/*---------------------------------------------------------------\
             SOME SETUP CONFIGURATION-UPDATE FOR EVERY YEAR PROCESSED
--------------------------------------------------------------------------*/
// Specify location of folder containing the current batch of NSF files.
// Also, specify the Fiscal Year being processed. NSF lets you download
// each FY's worth of awards files at a time.
$awardsFilePath = "C:/xampp/htdocs/thesis/WoS/1979/*.xml";
$awardsFY = 1979;
// Prep error file:
$errorFile = 'errors.txt';
file_put_contents($errorFile, "", FILE_APPEND);
/*-------------------------------------------------------------------------
 1.0 Open an active WoS session using functions included in
WoS_manage_sessions.php
--------------------------------------------------------------------------*/
$session = getSessionID();
echo("Session ID passed is: $session\n\n");
/*-------------------------------------------------------------------------
 2.0 Prepare the MySQL database connection.
-------------------------------------------------------------------------*/
$myDatabaseConnection=openDB();
/*-------------------------------------------------------------------------
 3.0 Parse Each File in NSF Awards folder
-------------------------------------------------------------------------*/
$filecount = 0;
foreach (glob($awardsFilePath) as $filename) {
$awardArray = parseAwardFile($filename, $awardsFY, $myDatabaseConnection);

/*-------------------------------------------------------------------------
 3.1 Prep & Execute MySQL INSERT Statement for Grant Record
--------------------------------------------------------------------------*/
$insertAwardQuery=convertAwardToSQLquery($awardArray);
executeSQL($insertAwardQuery,$myDatabaseConnection);
/*-------------------------------------------------------------------------
3.2 Execute a WoS Publication Search for Each Grant Record
--------------------------------------------------------------------------*/
$articlesResultSet=doSearch($session, $awardArray[AwardID],
$awardArray[AwardFY]);
$articlesResultSetXML=simplexml_load_string($articlesResultSet); /*Load
results contents into an XML object*/
/*Report any errors loading file contents to an XML object*/
foreach( libxml_get_errors() as $error ) {
print_r($error);
echo("\n");
file_put_contents($errorFile, "Load XML error:". $error . "\n",
FILE_APPEND);}

//Parse & organize desired elements. Return an array of strings containing
each article as an SQL insert statement
$arrayofArticles = parsePublicationRecords($articlesResultSetXML,
$myDatabaseConnection);
/*--------------------------------------------------------------------------
3.3 Convert Each Article into an SQL Statement and Insert into Database
---------------------------------------------------------------------------*/
$length=count($arrayofArticles);
//echo ("\n\nNumber of articles for Grant $awardArray[AwardID] is
$length\n\n");
for ($i=0;$i<$length;$i++){ /*for every article stored as a subarray*/
$sqlArticleInsert=convertArticleSubArrayToSQLinsert($arrayofArticles[$i],
$myDatabaseConnection);/*convert that article data to an insert SQL
statement*/
executeSQL($sqlArticleInsert,$myDatabaseConnection ); /*and execute the SQL
to insert the article into the database*/
/*next form another SQL statement to insert the connection between the award
and the article as a row in Award-To-Article table*/
$sqlA2Ainsert = createAward2ArticleAssociationSQL($awardArray[AwardID],
$arrayofArticles[$i][UID]);
executeSQL($sqlA2Ainsert, $myDatabaseConnection);}/*End for every article*/
}/*end for each award file*/
/*--------------------------------------------------------------------------
  After processing all NSF files and searching for their corresponding
publications,
  Close the WoS Session and Close the Database Connection
--------------------------------------------------------------------------*/
closeSession($session);
echo("Session closed.\n");
closeDB($myDatabaseConnection);
$time_end = microtime(true);
$execution_time = ($time_end - $time_start)/60;
echo("\n\n MINUTES ELAPSED: $execution_time\n\n")
?>
```

File Contents for database_connection.php

```
<?php

echo("database_connection.php included.\n");
/*****************************************************************************
**
*                          OPEN MYSQL CONNECTION
******************************************************************************
*/
function openDB(){
       $servername = "your_database_host";
       $username = "your_username";
       $password = "your_password";
       $dbname = "your_database";
       // Create connection
       $connection = new mysqli($servername, $username, $password, $dbname);
       // Check connection
       if ($connection->connect_error) {
       die("Connection failed: " . $connection->connect_error);}
       echo("\nDatabase opened.\n");
       return($connection);
}

/*****************************************************************************
**
*                          CLOSE  MYSQL CONNECTION
******************************************************************************
*/
function closeDB($connection){
$connection->close();
echo("\nDatabase closed.\n");
}
/*****************************************************************************
**
*                                 EXECUTE SQL QUERY
******************************************************************************
*/
function executeSQL($query, $connection){
       $errorFile = 'errors.txt';
       if ($connection->query($query) === TRUE) {
              echo "Record inserted.\n";}
              else {
              echo "Error: " . $query . "\n" . $connection->error;
              file_put_contents($errorFile, "Error: " . $query . "\n" .
$connection->error, FILE_APPEND);}
}
?>
```

File Contents for process_NSF_awards.php

```
<?php
echo("process_NSF_awards.php included.\n\n");
/* **************************************************************************
    File Name: process_NSF_award.php
       Author: Monica Inez Ihli
  Description: This file is called by thesis_main.php. It includes a function
               which receives a simple XML object containg a the contents of
                       an NSF award file. The NSF award files are downloaded
from
                       http://www.nsf.gov/awardsearch/download.jsp. The
function
                       returns the parsed values as an array which is
convenient for
                       further processing. It also includes a function for
building
               an SQL statement with the values
               for insertion into MySQL.

************************************************************************** */

function parseAwardFile($awardFileName, $awardFY, $myDatabaseConn)
{
       echo("Processesing: $awardFileName.\n");
       $inputFile = simplexml_load_file($awardFileName); /*Load file contents
into an XML object*/
       foreach( libxml_get_errors() as $error ) { /*For any errors loading
file to an XML object*/
             print_r($error);
             echo("\n");
       file_put_contents($errorFile, "Load XML error:". $error . "\n",
FILE_APPEND);}

       /*$investigatorsLNamesArray=array(); //Decided not to do name matching
against authors*/
       $investigatorNames="";

       foreach ($inputFile->Award->Investigator as $Investigator){
             /*$investigatorsLNamesArray[] = (string)$Investigator-
>LastName;*/
             $investigatorNames=$investigatorNames. (string)$Investigator-
>FirstName . ', ' .
                    (string)$Investigator->LastName . '; ';
       }
       /*The dates are parsed separately first so they can be converted before
       assignment into the array. This was the only way I could get dates to
parse correctly*/
       $AwardEffectiveDate=$inputFile->Award->AwardEffectiveDate;
       $AwardExpirationDate=$inputFile->Award->AwardExpirationDate;

       $NSFawardArray = array(
       'AwardID'=>(string)$inputFile->Award->AwardID,
       'AwardTitle'=>mysqli_real_escape_string($myDatabaseConn,
(string)$inputFile->Award->AwardTitle),
       'AwardAmount'=>(float)$inputFile->Award->AwardAmount,
       'AwardInstrument'=>(string)$inputFile->Award->AwardInstrument->Value,
       'OrganizationCode'=>(string)$inputFile->Award->Organization->Code,
       'DirectorateLongName'=>mysqli_real_escape_string($myDatabaseConn,
(string)$inputFile->Award->Organization->Directorate->LongName),
       'DivisionLongName'=>mysqli_real_escape_string($myDatabaseConn,
(string)$inputFile->Award->Organization->Division->LongName),
       'AbstractNarration'=> mysqli_real_escape_string($myDatabaseConn,
(string)$inputFile->Award->AbstractNarration),
       'InstitutionName'=>mysqli_real_escape_string($myDatabaseConn,
(string)$inputFile->Award->Institution->Name),
       'InstitutionCityName'=>mysqli_real_escape_string($myDatabaseConn,
(string)$inputFile->Award->Institution->CityName),
       'InstitutionStateCode'=>(string)$inputFile->Award->Institution-
>StateCode,
       'AwardEffectiveDate'=>date('Y-m-d', strtotime($AwardEffectiveDate)),
       'AwardExpirationDate'=>date('Y-m-d', strtotime($AwardExpirationDate)),
       'InvestigatorNames'=>mysqli_real_escape_string($myDatabaseConn,
$investigatorNames),
       'AwardFY'=>$awardFY,
       /*'investigatorsLNames'=>$investigatorsLNamesArray*/
       );
       return($NSFawardArray);
}

/* **************************************************************************
*/
function convertAwardToSQLquery($awardArray)
{
       $sql = "INSERT INTO AWARD(AwardID, AwardFY, AwardTitle,
AwardEffectiveDate, AwardExpirationDate,
       AwardAmount, AwardInstrument, OrganizationCode, DirectorateLongName,
DivisionLongName, AbstractNarration,
       InstitutionName, InstitutionCityName, InstitutionStateCode,
InvestigatorNames)
       VALUES ('$awardArray[AwardID]', '$awardArray[AwardFY]',
'$awardArray[AwardTitle]', '$awardArray[AwardEffectiveDate]',
       '$awardArray[AwardExpirationDate]', '$awardArray[AwardAmount]',
       '$awardArray[AwardInstrument]','$awardArray[OrganizationCode]',
'$awardArray[DirectorateLongName]',
       '$awardArray[DivisionLongName]', '$awardArray[AbstractNarration]',
       '$awardArray[InstitutionName]', '$awardArray[InstitutionCityName]',
'$awardArray[InstitutionStateCode]', '$awardArray[InvestigatorNames]')";
       return($sql);
}

?>
```

File Contents for WoS_manage_sessions.php

```
<?php
echo("WoS_manage_sessions.php included.\n");
/*
*****************************************************************************
******
             OPEN & RETURN A NEW WEB OF SCIENCE API SESSION

*****************************************************************************
******* */

function getSessionID()
{
$soap_post_string = "<soapenv:Envelope
xmlns:soapenv=\"http://schemas.xmlsoap.org/soap/envelope/\"
xmlns:auth=\"http://auth.cxf.wokmws.thomsonreuters.com\">
   <soapenv:Header/>
   <soapenv:Body>
      <auth:authenticate/>
   </soapenv:Body>
</soapenv:Envelope>";

//     Despite documentation, SOAPAction header should be excluded.
// "SOAPAction: [\"\"]",

$headers = array(
       "Content-Type: text/xml;charset=\"utf-8\"",
       "Accept: [*]",
       "connection=[keep-alive]",
       "host=[10.224.10.63:8081]",
       "Authorization=[Basic VVRLX0hHOndzUDZrbngh]",
       "Cache-Control: no-cache",
       "Pragma: no-cache",
       "Content-length: ".strlen($soap_post_string)
       );

$curl = curl_init();
curl_setopt_array($curl, array(
    CURLOPT_RETURNTRANSFER => 1,
    CURLOPT_URL =>
'http://search.webofknowledge.com/esti/wokmws/ws/WOKMWSAuthenticate/auth:auth
enticate',
       CURLOPT_TIMEOUT => 10,
    CURLOPT_POST => 1,
       CURLOPT_POSTFIELDS => $soap_post_string,
       CURLOPT_HTTPHEADER => $headers));

$sessionID = curl_exec($curl);
curl_close($curl);
echo ("\n\n\n");
//Because "return" is a reserved keyword in php, I can't compile any
reference to the element
//So I am using a string replace to substitute some other reference to the
element.
//These next few lines are stripping out the soap elements. I just want the
Session ID.
$sessionID = str_replace("<soap:Envelope
xmlns:soap=\"http://schemas.xmlsoap.org/soap/envelope/\"><soap:Body><ns2:auth
enticateResponse xmlns:ns2=\"http://auth.cxf.wokmws.thomsonreuters.com\">",
"", $sessionID);
$sessionID = str_replace("<return>", "", $sessionID);
$sessionID =
str_replace("</return></ns2:authenticateResponse></soap:Body></soap:Envelope>
", "", $sessionID);
return $sessionID;
}

/*****************************************************************************
*****
                           CLOSE WEB OF SCIENCE API SESSION
******************************************************************************
*** */
function closeSession($sessionToClose){
/*The soap message will be stored in a string. The string will later be
passed as the post data*/
/*Note that the session ID will be sent in the HTTP header and is not part of
the soap message. -->*/

$soap_post_string = "<soap:Envelope
xmlns:soap=\"http://schemas.xmlsoap.org/soap/envelope/\">
   <soap:Body>
   <WOKMWSAuthentcate:closeSession
   xmlns:WOKMWSAuthentcate=\"http://auth.cxf.wokmws.thomsonreuters.com\"/>
   </soap:Body>
   </soap:Envelope>";


/*Construct an array containing the various headers we will pass for the
header parameter of the cURL request.*/
$headers = array(
       "Content-Type: text/xml;charset=\"utf-8\"",
       "Accept: [*]",
       "Cookie: SID=\"$sessionToClose\"", /*input the Session ID returned from
the authorization */
       "Cache-Control: no-cache",
       "Pragma: no-cache",
       "Content-length: ".strlen($soap_post_string)
       );


$curl = curl_init();
curl_setopt_array($curl, array(
    CURLOPT_RETURNTRANSFER => 1,
    CURLOPT_URL =>
'http://search.webofknowledge.com/esti/wokmws/ws/WOKMWSAuthenticate/auth:auth
enticate',
       CURLOPT_TIMEOUT => 10,
    CURLOPT_POST => 1,
       CURLOPT_POSTFIELDS => $soap_post_string,
       CURLOPT_HTTPHEADER => $headers));

$closeResponse = curl_exec($curl);
curl_close($curl);
}
?>
```

File Contents for WoS_execute_search.php

```
<?php
echo("WoS_execute_search.php included.\n");
/*
******************************************************************************
******
                                        ACCEPT SESSION ID & SEARCH
PARAMETERS, EXECUTE
                                                      QUERY, AND RETURN THE
RESULTING DATA

******************************************************************************
******* */

function doSearch($searchSession, $awardID, $awardFY)
{
       /* removed from below collection. This is optional. Not including it
means all WoS dB in subscription
       will be searched.
       <editions>
       <collection>WOS</collection>
       <edition>SCI</edition>
       </editions>
       */
       // Use this parameter for begin if want to limit results to those after
the FY
       //<begin>".$awardFY."-01-01</begin>
       // This is where the query is formed, within <userQuery>
       $soap_post_string =
       "<soapenv:Envelope
xmlns:soapenv=\"http://schemas.xmlsoap.org/soap/envelope/\"
       xmlns:woksearch=\"http://woksearch.v3.wokmws.thomsonreuters.com\">
       <soapenv:Header/>
       <soapenv:Body>
       <woksearch:search>
       <queryParameters>
       <databaseId>WOS</databaseId>
       <userQuery>FO=(NSF OR National Science Foundation) AND
FG=(".$awardID.")</userQuery>
       <timeSpan>
              <begin>2010-01-01</begin>
              <end>2014-12-31</end>
       </timeSpan>
       <queryLanguage>en</queryLanguage>
       </queryParameters>
       <retrieveParameters>
       <firstRecord>1</firstRecord>
       <count>100</count>
       </retrieveParameters>
       </woksearch:search>
       </soapenv:Body>
       </soapenv:Envelope>
       "; /*count is where you can specify max records to return*/

       /*Construct an array containing the various headers we will pass for
the header parameter of the cURL request.*/
       $headers = array(
              "Content-Type: text/xml;charset=\"utf-8\"",
              "Accept: [*]",
                "Cookie: SID=\"$searchSession\"", /*input the Session ID
returned from the authorization */
                "Cache-Control: no-cache",
                "Pragma: no-cache",
                "Content-length: ".strlen($soap_post_string)
                );


        $curl = curl_init();
        curl_setopt_array($curl, array(
                CURLOPT_RETURNTRANSFER => 1,
                CURLOPT_URL =>
'http://search.webofknowledge.com/esti/wokmws/ws/WokSearch/woksearch:search',
                CURLOPT_TIMEOUT => 10,
                CURLOPT_POST => 1,
                CURLOPT_POSTFIELDS => $soap_post_string,
                CURLOPT_HTTPHEADER => $headers));

        $response = curl_exec($curl);
        curl_close($curl);
        //strip the soap message elements away, leaving just the XML data.
        $response=str_replace("<soap:Envelope
xmlns:soap=\"http://schemas.xmlsoap.org/soap/envelope/\"><soap:Body><ns2:sear
chResponse
xmlns:ns2=\"http://woksearch.v3.wokmws.thomsonreuters.com\">","",$response);
        $response=str_replace("</ns2:searchResponse></soap:Body></soap:Envelope
>","", $response);

        //clean up by replace &lt; and &gt; with appropriate symbols
        $response=str_replace("&lt;","<", $response);
        $response=str_replace("&gt;",">", $response);
        //return is reserve word so replace the element name with something I
can work with.
        $response = str_replace("<return>", "<r3turn>", $response);
        $response = str_replace("</return>", "</r3turn>", $response);

        return($response);

}

/* This function receives a simpleXMLobject containing the set of publication
records
returned from a WoS search, and converts each publication into an array of
values for the article.
It also accepts the database connection as a parameters so that it can
perform mysqli_escape_string()
on the data which will be inserted
The article -level arrays become subarrays when assigned to a main array
which serves as a container for passing all the
articles back to the main program.
 */

function parsePublicationRecords($articleResultSetXML, $myDatabaseConn){
        $articlesArray=array();
        foreach ($articleResultSetXML->records->records->REC as $rec): /*For
each publication in the result set*/
        /*First we have to deal with parsing each of the titles*/
        foreach ($rec->static_data->summary->titles->title as $title): /*For
every title element in a record*/
```

```
       /*there are six title elements returned: 5 variants of the journal
title and the final title type is the title of the publication itself*/
       switch((string) $title['type']){
       case 'source':
       $source_title = $title;
       break;
       case 'source_abbrev':
       $source_abbrev = $title;
       break;
       case 'abbrev_iso':
       $abbrev_iso = $title;
       break;
       case 'abbrev_11':
       $abbrev_11 = $title;
       break;
       case 'abbrev_29':
       $abbrev_29 = $title;
       break;
       case 'item':
       $item_title = $title;
       break;
       }
       endforeach; /*for each title in pub*/

       // Static Data > Summary
       $UID = (string)$rec->UID;
       $pubtype= (string)$rec->static_data->summary->pub_info['pubtype'];
       $pubmonth= (string)$rec->static_data->summary->pub_info['pubmonth'];
       $vol= (string)$rec->static_data->summary->pub_info['vol'];
       $pubyear= (int)$rec->static_data->summary->pub_info['pubyear'];

       // static data > fullrecord_metadata
       $fund_text = (string)$rec->static_data->fullrecord_metadata->fund_ack-
>fund_text->p;

       /*For convenience to store all the grant info in a single field, purely
to support
       record-level human analysis if I want to examine any on a case-by-case
basis*/
       /*grants are actually organized as one grant field for each agency,
with multiple grant IDs for more than one grant per agency/grant*/
       $grantsString="";
       foreach ($rec->static_data->fullrecord_metadata->fund_ack->grants-
>grant as $grant){
       $grantsString = $grantsString . (string)$grant->grant_agency . ': ';
       foreach ($grant->grant_ids->grant_id as $grant_id){
       $grantsString = $grantsString . $grant_id . ', ';
       } //end for each grant_id
       $grantsString = $grantsString.'; ';
       }//end for each grant


       $oneArticle=array(
       'UID' =>$UID,
       'source_title'=>$source_title,
       'source_abbrev'=>$source_abbrev,
       'abbrev_iso'=>$abbrev_iso,
       'abbrev_11'=>$abbrev_11,
       'abbrev_29'=>$abbrev_29,
       'item_title'=>$item_title,
       'pubtype'=>$pubtype,
       'pubmonth'=>$pubmonth,
       'vol'=>$vol,
       'pubyear'=>$pubyear,
       'fund_text'=>mysqli_real_escape_string($myDatabaseConn, $fund_text),
       'grants'=>mysqli_real_escape_string($myDatabaseConn, $grantsString)
       );
       $articlesArray[]=$oneArticle;

       endforeach;
       return ($articlesArray);

}
/*********************************************************************/
function convertArticleSubArrayToSQLinsert($articleSubarrayElement)
{
       $sqlString = "INSERT INTO ARTICLE(UID, source_title, source_abbrev,
       abbrev_iso, abbrev_11, abbrev_29,
       item_title, pubtype, pubmonth,
       vol, pubyear, fund_text, grants)
       VALUES ('$articleSubarrayElement[UID]',
'$articleSubarrayElement[source_title]',
'$articleSubarrayElement[source_abbrev]',
       '$articleSubarrayElement[abbrev_iso]',
       '$articleSubarrayElement[abbrev_11]',
'$articleSubarrayElement[abbrev_29]',
       '$articleSubarrayElement[item_title]','$articleSubarrayElement[pubtype]
', '$articleSubarrayElement[pubmonth]',
       '$articleSubarrayElement[vol]', '$articleSubarrayElement[pubyear]',
       '$articleSubarrayElement[fund_text]',
'$articleSubarrayElement[grants]')";

       return($sqlString);

}

/*********************************************************************/
function createAward2ArticleAssociationSQL($awardID, $UID)
{

       $sqlString = "INSERT INTO AWARDTOARTICLE(AwardID, UID)
       VALUES ('$awardID', '$UID')";

       return($sqlString);
}



?>
```

Vita

Monica Ihli was born in Lake Charles, Louisiana. She graduated with an Associate of Applied Science in Computer Sciences from Pellissippi State Community College in 2005, and graduated with Bachelor of Art in Communication Studies from the University of Tennessee at Knoxville in 2013. She has contracted as a systems analyst with Oak Ridge National Laboratory, has interned with the National Aeronautics and Space Administration while studying the structure and evolution of a scientific community of practice, and has been employed with the University of Tennessee Libraries managing enterprise-level technology systems. In August, 2014, she entered the Graduate School of the University of Tennessee at Knoxville, in the College of Communication and Information’s Information Sciences program.