Scientometrics

Assessing Researcher Interdisciplinarity: A Case Study of the University of Hawaii NASA Astrobiology Institute
--Manuscript Draft--

Manuscript Number:	SCIM1640R2
Full Title:	Assessing Researcher Interdisciplinarity: A Case Study of the University of Hawaii NASA Astrobiology Institute
Article Type:	Manuscript
Keywords:	Astrobiology; Bibliometrics; Information bottleneck method; Interdisciplinary science; Machine learning; Text mining
Corresponding Author:	Michael Gowanlock University of Hawaii Honolulu, HI UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Hawaii
Corresponding Author's Secondary Institution:	
First Author:	Michael Gowanlock
First Author Secondary Information:	
Order of Authors:	Michael Gowanlock
	Rich Gazan
Order of Authors Secondary Information:	
Abstract:	In this study, we combine bibliometric techniques with a machine learning algorithm, the sequential Information Bottleneck, to assess the interdisciplinarity of research produced by the University of Hawaii NASA Astrobiology Institute (UHNAI). In particular, we cluster abstract data to evaluate Thomson Reuters Web of Knowledge subject categories as descriptive labels for astrobiology documents, assess individual researcher interdisciplinarity, and determine where collaboration opportunities might occur. We find that the majority of the UHNAI team is engaged in interdisciplinary research, and suggest that our method could be applied to additional NASA Astrobiology Institute teams in particular, or other interdisciplinary research teams more broadly, to identify and facilitate collaboration opportunities.
Response to Reviewers:	We have separated the Appendix as an Online Supplement. Also, we elected to move one of the tables from our previous manuscript submission from the appendix to the main text. This table appears as Table 1 in the revised manuscript.

Assessing Researcher Interdisciplinarity: A Case Study of the University of Hawaii NASA Astrobiology Institute

Michael Gowanlock & Rich Gazan

M. Gowanlock

Department of Information & Computer Sciences and University of Hawaii NASA Astrobiology Institute, University of Hawaii

POST 310, 1680 East-West Road, Honolulu, HI 96822, USA

gowanloc@hawaii.edu

R. Gazan

Department of Information & Computer Sciences, Library & Information Science Program and University of Hawaii NASA Astrobiology Institute, University of Hawaii, Hamilton Library 2H, 2550 McCarthy Mall, Honolulu, HI 96822, USA

gazan@hawaii.edu

Abstract

In this study, we combine bibliometric techniques with a machine learning algorithm, the sequential Information Bottleneck, to assess the interdisciplinarity of research produced by the University of Hawaii NASA Astrobiology Institute (UHNAI). In particular, we cluster abstract data to evaluate Thomson Reuters Web of Knowledge subject categories as descriptive labels for astrobiology documents, assess individual researcher interdisciplinarity, and determine where collaboration opportunities might occur. We find that the majority of the UHNAI team is engaged in interdisciplinary research, and suggest that our method could be applied to additional NASA Astrobiology Institute teams in particular, or other interdisciplinary research teams more broadly, to identify and facilitate collaboration opportunities.

Keywords: **Astrobiology, Bibliometrics, Information bottleneck method, Interdisciplinary science, Machine learning, Text mining**

1   Introduction

Astrobiology, the study of the origin, evolution, distribution, and future of life in the universe, is a relatively new field comprised of researchers from a range of scientific disciplines. Apart from its sublime object of study, astrobiology has been identified as a field that can integrate diverse sciences (Staley, 2003), provide a tangible target for interdisciplinary science education (Cockell, 2002), and provide a pathway to adult science literacy (Oliver and Fergusson, 2007). Many of the field's core questions require knowledge from multiple disciplines to be harvested, integrated and applied outside of their source domains, and as such, astrobiology is inherently interdisciplinary. For example, the University of Hawaii NASA Astrobiology Institute (UHNAI) studies the origin of water in the solar system and beyond, in the context of understanding the origins of life. Astronomers, chemists, geologists, oceanographers and biologists work together to study data from meteorite fragments to comets to the interstellar medium to address the question of where else in the universe water, and thus life, might be found. Without collaboration across disciplinary boundaries to interpret often-scarce data, important questions in astrobiology will remain incompletely addressed. Developing a method to identify, measure and catalyze interdisciplinary work in the astrobiology research environment is the goal of this paper.

One of the benefits of a broad-based research community is that new developments in astrobiology occur fairly frequently. The downside is that researchers must stay abreast of these numerous developments both inside and outside of their home fields. As new astrobiology research findings are reported, the considerable effort involved in finding, evaluating and integrating them indicates a need for a better understanding of how findings in one field might inform others, and to identify potential collaboration opportunities between individual researchers working on similar questions.

Our previous example suggested that knowledge from multiple disciplines is required to understand the origin of water to answer questions regarding the origin of life. Satisfactorily understanding the research record of scientists that work in this area requires measuring interdisciplinarity on an acute scale. Following van Leeuwen (2007), we distinguish between a top-down bibliometric approach, where large-scale trends at the highest levels of publication aggregation are considered (such as the research output of a country or university), and prefer a bottom-up approach, where we analyze individual documents and the papers they cite. We harvest each astrobiologist's publication data by comparing NASA Astrobiology Institute annual reports, where publications are systematically documented, with the researchers' Websites and CVs, and further verify the data by browsing the author indexes of each database to identify name variations, to represent the research output of each astrobiologist.

A common method used to examine the potential of collaboration across disciplinary boundaries is to interview domain experts, but this method suffers from several limitations, such as sample size and subjectivity problems (Zhang et al., 2011). Furthermore, given that the subject matter of astrobiology spans many disciplines, meaningful analysis of the responses would require the knowledge of an astrobiology polymath. After considering these limitations, we suggest that measuring interdisciplinarity should be guided by one or more individuals versed in astrobiology, but whose expertise need not span all of its constituent disciplines. Therefore, an unsupervised approach is optimal as such methods can find trends in data without prior knowledge of its structure.

As of 2011, the NASA Astrobiology Institute is comprised of 14 teams spanning ten universities in addition to NASA Ames, Goddard, and the Jet Propulsion Laboratory. While a cross-team analysis is beyond the scope of this paper, we suggest that our method for measuring researcher interdisciplinarity at UHNAI could be extended to other NASA Astrobiology Institute teams, and to scientific collaborations more broadly. Furthermore, our method suggests where collaborations

might productively occur, and allows us to better understand the nature of interdisciplinary scientific discovery.

In this pilot study, we investigate the use of an unsupervised machine learning clustering technique, the sequential Information Bottleneck (sIB) (Slonim et al., 2002) to aid in measuring researcher interdisciplinarity. Furthermore, we assess the extent to which Journal Subject Categories from the Thomson Reuters Web of Knowledge database suite are sufficient for labelling astrobiology documents. The clustering and classification of text allow interdisciplinary analysis that 1) describes collaboration and the integration of knowledge and 2) draws conclusions that are useful to astrobiology researchers by uncovering the underlying structure of research tracks. The results of this pilot study will serve to guide a subsequent investigation that will identify collaboration opportunities and measure the disciplinary roots across the entire NASA Astrobiology Institute.

Researchers in astrobiology tend to be comfortable speaking in the language of multiple scientific disciplines. As suggested in Gargaud and Tirard (2011), these interdisciplinary researchers are somewhat isolated from their counterparts in other academic departments. The multidisciplinary context given by astrobiology affords an excellent opportunity to examine the methods used to study researcher interdisciplinarity and knowledge integration. Furthermore, we propose an iterative process to identify specific publications that bridge diverse fields, to facilitate interdisciplinary collaborations and ease the cognitive load of a single researcher who wishes to integrate knowledge from multiple disciplines.

2 Background

Research that occurs at the intersection between disciplines is thought to lead to great advances in science (Porter and Rafols, 2009). Many funding agencies exist specifically to support and encourage interdisciplinary research; the U.S. National Science Foundation's interdisciplinary research efforts span all of their divisions and directorates (National Science Foundation, Accessed November 21, 2011). For example, some authors measuring interdisciplinarity lament that there is not enough coverage of the societal causes for climate change (Bjurström and Polk, 2011) as described in the Intergovernmental Panel on Climate Change (IPCC) literature. In this specific case, measuring both the disciplinary diversity and the integration of knowledge is of paramount importance to

ensure that future IPCC reports include appropriate factors. A cynical disposition to this problem is eloquently stated in Brewer (1999): "The world has problems, but universities have departments."

Many important terms in this work have so far been discussed without qualification. The term interdisciplinary tends to be tacitly understood by researchers, with no consensus definition. We adopt the definition suggested by Porter et al. (2007), which followed the definition given by the National Academies (2005): interdisciplinary research requires an integration of concepts, theories, techniques and/or data from two or more bodies of specialized knowledge. Multidisciplinary research may incorporate elements of other bodies of specialized knowledge, but without interdisciplinary synthesis (Wagner et al., 2011) that leads to research that is greater than the sum of its parts.

Despite the increase in claimed interdisciplinarity, traditional indicators are of questionable value in assessing and quantifying interdisciplinary research (Morillo et al., 2001). Additionally, policies regarding interdisciplinarity are often based more on conventional wisdom than empirical studies (Rafols and Meyer, 2010). The usefulness of bibliometric indicators depends critically on the level at which we wish to understand the integrative process. For example, funding agencies may only require high-level publication co-authorship and collaboration statistics, describing the research performed by their grantees and the diversity of their home disciplines, but not addressing the essential aspect of synthesis. When there is no mechanism to identify, measure and encourage these points of intellectual crossover, there is no way to quantify the extent to which interdisciplinary science is taking place.

Top-down approaches have been used to map scientific literature (for example, see Boyack et al. (2005)), and often represent broad areas of science with Web of Knowledge (WoK) subject categories (SCs). For example, van Raan and van Leeuwen (2002) and Porter et al. (2007) used SCs in their methodology to measure interdisciplinarity. In these studies, SCs have been employed as de facto disciplinary boundaries, and as a benchmark to measure how much a given author, journal or research area crosses scientific fields. Unfortunately, low-level conclusions that might inform potentially productive individual collaborations cannot be made when relying on these top-down approaches, as they focus on past outputs rather than future integration. Conversely, bottom-up bibliometric approaches incorporate the authors' own words, in free-text fields such as: titles,

abstracts, keywords[1] and the full text of a document. Clustering bibliometric data at this level can describe the structure of a researcher, journal or an entire field, and suggest productive future directions. Comparing the bottom-up clustered output with the top-down approach of SCs for astrobiology publications yields an indication of the effectiveness of SCs as document labels for works in this interdisciplinary domain. Kostoff (1998) describes how a citation analysis can serve as a "radioactive trace" of research impacts. One limitation of cluster analysis is that "...precise disciplinary divisions are not obtained, rendering inter-cluster links misleading" (Small, 2010), but Upham and Small (2010) propose a methodology to identify emerging "research fronts", highly cited micro-specialty areas that transcend existing fields. Their method requires that the researcher not presuppose the existence of any research field, but to rely instead on a comprehensive monitoring of citations to identify points across disciplines where research interests intersect, echoing one goal of the present study. Both top-down and bottom-up approaches are useful in different applications. A study by Rafols and Meyer (2010) combines bottom-up and top-down approches to measure both disciplinary diversity and knowledge integration.

Measuring scientific output in bibliometric terms requires some degree of integration and normalization of the publication records of researchers, which are published in diverse formats, venues and scholarly traditions. The publication record generally includes data such as departmental affiliations, keywords, year of publication, journal, cited references, and the abstract and/or full text of the publication. This data can be compared using various bibliometric techniques to assess interdisciplinary research. While bibliometric studies tend to rely on a citation analysis, such an analysis may not be appropriate for every discipline or field. For example, a given field may tend to reference conference proceedings, websites, newspapers, or colloquia which are not as conducive to a co-citation analysis as journal articles. Due to this observation, Sugimoto (2011) suggests that studying interdisciplinarity should include publications beyond journal articles. While we agree with this position, it happens that journals are the preferred method of communication within the great majority of the fields that compose the UHNAI team; therefore, the present study is not hindered by this limitation.

---

[1]Keywords are not always a free-text field.

3   Methodology

In this section, we outline our method for measuring interdisciplinary research. In the previous section we noted that particular bibliometric indicators are conducive to understanding research at varying levels. One of the goals of this research is to uncover the underlying structure within an astrobiology research team that undertakes interdisciplinary projects at the macro scale, but may differ in the extent of interdisciplinary work at the micro level. To understand the research structure, we examine the abstract text of research publications and employ a method from the field of information theory, the sIB method, to cluster our high dimensional abstract data.

An advantage of using WoK for bibliometric studies is that it provides a mapping of SCs to each journal. Given the incommensurability of other bibliometric data (for example, journals do not agree upon a common set of keywords), SCs provide a way to compare publications on the journal level. In Zhang et al. (2010), the authors used a cross-citation analysis to create seven high-level clusters of related SCs, though their analysis was somewhat confounded by the "idiosyncrasy" that a journal may be assigned to multiple SCs in WoK. In Porter et al. (2007), the authors examine the references in sets of journal articles gathered from WoK, and relate the journals to their corresponding SCs. In this approach, a more diverse set of SCs that represent a paper derived from its references indicates a higher degree of interdisciplinarity than a set of similar SCs that represent a paper.

Using the references of a paper is a reasonable approach to measure researcher interdisciplinarity. Analoguous to Porter et al. (2007), we use the references in each UHNAI publication. In particular, we combine all of the abstracts of all of the references cited by a UHNAI publication, and use these aggregated abstracts to represent each publication. In another text mining study (Kostoff et al., 2001), employed free-text fields (such as title, keywords and abstracts) of cited/citing publications in combination with phrase frequency analysis and phrase clustering analysis to obtain a low-level understanding of research impact and interdisciplinary research.

In the present study, we focus on the abstracts of cited papers only, and we do not consider the papers that cite the UHNAI papers. A major limitation of studying the citations to the UHNAI papers is that it would require the database to contain those papers that cite a particular work, which varies between disciplines, fields and databases. The same is true of those references that

6

are cited in our UHNAI papers. To obviate this problem, we elect to use the NASA Astrophysics Data System (ADS) to collect the majority of our abstracts, as more UHNAI team publications are covered in this database than any other. The extensive coverage in ADS ensures that a considerable majority of papers referenced by the UHNAI team are also within the database. However, previous research has illustrated how the differences in scientific publication patterns between fields often require that records from multiple databases be harvested to encompass the output of interdisciplinary scientific researchers (see, for example, Kousha and Thelwall, 2008). For UHNAI authors whose publications were not sufficiently represented in ADS, we used WoK to obtain their publication data and cited references. As it turns out, those authors, and the papers they cite, were highly represented in WoK. We were able to gauge author coverage in ADS and WoK by consulting the CV of each UHNAI team member.

In the following subsections, we describe our methods used to achieve the following goals:

- Examine whether WoK SCs are sufficient for labelling astrobiology documents. While we believe SCs are useful in mapping broad scientific research trends in established disciplines, whether they are appropriate for classifying individual publications or interdisciplinary works remains an open question[2]. We cluster a corpus of astrobiology abstracts labelled with their corresponding conflated SCs (Section 3.3), and assume that if a given cluster is comprised mostly of a single SC, then SCs are a sufficiently accurate classifier.

- Identify actual and potential instances of interdisciplinary research in astrobiology using conflated SCs (Section 3.3).

- Identify actual and potential instances of interdisciplinary research and identify potential collaboration opportunities between researchers using aggregated abstracts to represent the research tracks of the UHNAI team (Section 3.4).

3.1 Text Mining and the Sequential Information Bottleneck (sIB) Method

The sIB clustering algorithm (Slonim et al., 2002) is employed to cluster our datasets described below. We chose this clustering method over others because it has been shown to perform better

---

[2]The classification of documents is a requirement for an astrobiology publication information retrieval system. Our research group is inclined to create such a system. See `http://airframe.ics.hawaii.edu/` for more information.

than other unsupervised clustering methods, such as k-means (Slonim et al., 2002). Furthermore, the approach should allow us to identify instances of interdisciplinary research by examining the cluster membership of our abstract data without prior knowledge of the data's properties. It is necessary to use an unsupervised clustering method because a canonical set of astrobiology documents with which to train a clustering technique does not exist.

3.2 Data Collection

We gather publications by the UHNAI team members from 2001 until June 2011. Publications earlier than 2001 were not collected, as many researchers may not have been engaged in astrobiology research, and the UHNAI team was not yet founded[3]. However, we place no age restrictions on the papers that they cite.

3.2.1 NASA Astrophysics Data System (ADS)

The ADS has extensive coverage of astronomy, astrophysics and physics journal articles, pre-prints and conference proceedings. We gather the data in a semi-automated fashion. Instead of accessing the articles through a web browser, ADS has a perl script library[4] that can be used to access parts of the database. To gather the abstracts and journals of UHNAI papers, and the abstracts and journals of the papers they cite, we employed the following procedure:

- Ran one of the ADS perl scripts to return a list of all of the publications for each UHNAI team member. This returned a list of ADS bibcodes, which uniquely identify each record in the ADS.

- Compared these papers with each author's CV to ensure that we did not collect undesired articles. For example, we filtered out papers by authors with the same last name as members on the UHNAI team.

- Used the ADS bibcodes to create a script that goes to the URL of the webpage that lists the references in each UHNAI paper. We download the individual webpages.

---

[3]This is also the year that the journal Astrobiology began publication. While astrobiology research was, and continues to be published in other journals, this indicates that astrobiology research may not have coalesced as a field prior to 2001.

[4]The scripts can be found here: `http://vo.ads.harvard.edu/adswww-lib/`

- Created and ran a script to capture all of the ADS bibcodes in each downloaded html webpage.

- Used this list of bibcodes to get the abstracts and journals of all of the UHNAI papers and references therein using the ADS perl scripts.

3.2.2 Web of Knowledge (WoK)

To include the published output of UHNAI authors whose work is underrepresented in ADS, and to provide a comprehensive portrait of the entire UHNAI team, we also used WoK to gather abstracts and bibliographic data. To our knowledge there is not an API or alternative way to access WoK other than using a web browser. To gather this data, we employed the following procedure:

- Created a list of all of the papers authored by UHNAI authors that were not in or underrepresented in the ADS database.

- Manually downloaded the html pages of each record describing each UHNAI publication and references therein.

- Created a script that parses the html pages to harvest the abstracts and journal titles.

When working with WoK data, it is important to be mindful of the differences in institutional subscriptions, which include access to different subsets and date ranges of WoK's constituent databases, and may affect the results of bibliometric analysis (Derrick et al., 2010; Jacsó, 2005). Therefore, we provide a list of the University of Hawaii WoK subscriptions at the time of data collection:

- Web of Science, 1980 -

- Biological Abstracts, 1969 -

- Medline, 1950 -

- Journal Citation Reports Science & Social Science editions, 2004 -

3.3 WoK Subject Categories and Document Classification

Having collected the abstracts and journal names of UHNAI publications and references, we create a dataset that contains the abstract text and the SC of the associated journal for each UHNAI publication and the publications they cite.

Many of the SCs of the papers in our dataset were significantly underrepresented. Furthermore, as other researchers have encountered (see, for example, Zhang et al., 2010) some journals in WoK are assigned multiple SCs, necessitating some conflation into superclusters, or "macro-disciplines" (Porter and Rafols, 2009). We modify the SCs using the following method:

- Journals with a single WoK SC that appears 10 or more times in our dataset uses the assigned WoK SC name.

- Journals with a single WoK SC that appears less than 10 times is changed to a broader WoK category (e.g. "Biochemical Research Methods" becomes "Biochemistry & Molecular Biology").

- Journals with two or more SCs of roughly equivalent weight are assigned a new conflated SC (e.g. "Astrophysics & Geophysics").

- Journals with two or more SCs that have a clear primary SC have "-Multidisciplinary" appended to the primary name.

The ADS system also contains non-journal publications. In these instances, we manually assigned an appropriate SC to the publication. Table 1 shows the mapping of WoK SCs to our conflated SCs.

We eliminated those abstracts whose SCs were unique or constituted a very small fraction of the entire dataset. Furthermore, publications are commonly cited across the UHNAI team; therefore, multiple duplicate abstracts could appear in our dataset. Duplicate abstracts were removed from the sample. Once removing the duplicate abstracts, we traced back which databases these abstracts came from to reflect the relative proportion of abstracts obtained through WoK and ADS. The dataset has 10216 abstracts integrated over 13 conflated SCs. Table 2 shows the number and fraction of abstracts labeled with each conflated SC. From Table 2, we observe that there is a large

class imbalance problem, as the Astronomy & Astrophysics SC contributes 67.68% of the entire dataset.

We oversample the minority classes (where a class is a SC), which is every SC but Astronomy & Astrophysics, to examine if the class imbalance problem significantly affects the resultant clusters. There are a number of methods utilized to oversample minority classes in the field of data mining. Duplicating the abstracts in the minority classes could potentially result in model overfitting. To obviate this problem, we create synthetic data that is similar to the other abstracts within a given SC. We use the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) to produce synthetic feature vectors, where a feature vector (or feature) is a normalized numerical representation of the words that describe each abstract/instance. For example, consider the following two truncated abstracts:

1. Water is found on the earth and in the solar system.

2. Water exists on the moon, and Mars.

The two feature vectors of word counts after punctuation is removed is shown in Figure 1.

For clustering our abstracts and their corresponding conflated SCs, we create two datasets. In the first dataset (hereafter *conflated_SC_default*), we cluster the dataset as described by Table 2, without considering the class imbalance problem. In the second dataset (hereafter *conflated_SC_sampled*), we randomly sample without replacement 25% of the features contained within the Astronomy & Astrophysics SC and every feature in the minority SCs three times. We use SMOTE to create synthetic feature vectors for the minority SCs such that each SC is represented by the same number of features. A visual representation of the distribution of real and synthetic data is shown in Figure 2.

3.4 Text Mining Aggregated Abstracts

We create a dataset of aggregated abstracts (hereafter *aggregated_abstracts*) for the purposes of representing each UHNAI publication. The dataset contains 731 publications by the UHNAI team. Table 3 shows the team members and their associated home disciplines. Each publication is represented by its own abstract and the abstract of each cited publication. We aggregate all of these

abstracts in a single feature vector to represent each UHNAI publication. Non-journal publications such as book chapters, conference proceedings and dissertations were included in the dataset, although they constitute a very small fraction of the total publications. A majority of the abstracts in the *aggregated_abstracts* dataset are the same as the ones in the *conflated_SC_default* and *conflated_SC_sampled* datasets.

We estimate the completeness of the aggregated abstracts, which is defined as the fraction of abstracts harvested out of the total number of citations in a UHNAI publication. For example, if an individual UHNAI publication contains 20 referenced citations, and 15 corresponding cited abstracts were harvested, then the aggregated abstract is 75% complete. We randomly sampled (N=100) abstracts from the 731 in the *aggregated_abstracts* dataset. We find that the average completeness for the aggregated abstracts in this sample is 74.3%, as shown in Table 4. Therefore, we expect that on average our aggregated abstracts in the *aggregated_abstracts* dataset are $\sim$74% complete. Interestingly, the UHNAI publications harvested from ADS have a higher degree of completeness than those abstracts harvested from WoK.

3.5 Preprocessing of the Datasets

We preprocessed the *conflated_SC_default*, *conflated_SC_sampled* and *aggregated_abstracts* datasets in the same manner. Our preprocessing of the datasets included converting uppercase words to lowercase, and ignoring non-alphabetical characters. We stemmed the words using the Porter stemming algorithm (Porter, 1980) to ensure that related words were not duplicated in the datasets. We created a stopword list to remove formatting tags, and other non-content-bearing terms. We selected words which had a minimum frequency of 12, integrated over the entire datasets, resulting in a total of $\sim$4000 words in each dataset. Most of our preprocessing was performed in WEKA (Witten and Frank, 2005), and the sIB method was also executed in this environment.

We normalize each feature vector in our datasets. Each feature is described by the term frequency of each word found in the $\sim$4000 words distilled from their respective datasets. We normalize the sum of each feature vector to 1. In the case of the aggregated abstracts, some feature vectors will be much shorter or longer than others, as there is a large range of abstract sizes, and number of references within a given publication. If we did not normalize the term frequencies, then instances with high word or low word counts may cluster together. Such clusters would be less revealing of

12

the content of the documents themselves. Figure 3 reiterates the steps employed to construct our datasets.

3.6   Limitations

There are several limitations to our study. First, some of the papers were authored by multiple members of the UHNAI team. In this case, we assigned the abstract data to the first-listed author on the paper, thereby not fully characterizing the research contribution of the non-primary authors. Otherwise, having multiple labels on the same document would inadvertently oversample those documents with multiple UHNAI authors. Also, there is a minor discrepancy between the abstracts gathered in ADS and WoK; ADS contains abstracts from non-journal sources, whereas WoK does to a lesser extent for the researchers studied here. The vast majority of our data was from journal articles; therefore, we do not expect this to have a significant, if any negative impact on our study. WoK maps multiple SCs to a single journal. While we need to conflate the SCs in order to compare them to clusters (in the *conflated_SC_default* and *conflated_SC_sampled* datasets), the aggregation procedure undermines the fundamental function of SCs. Furthermore, we reduced the total number of conflated SCs to 13 which may have a negative effect on our ability to assess interdisciplinary research.

4   Results

In this section we present the results of our text mining experiments. For the purposes of this paper, where our goal is to identify actual and potential instances of interdisciplinary research in astrobiology, a meaningful cluster relationship is one where papers from two or more SCs cluster together, or when researchers from different fields have the aggregated abstracts of their papers cluster together. Our present and future work is focused on these heterogeneous clusters, however our method could be used for a variety of purposes, each with a different corresponding indicator of interest. For example, a research team wishing to demonstrate its uniqueness within a collaboration might highlight its work being represented as a relatively homogenous cluster, with its dominant SC not found in other clusters. A group seeking to align or connect itself with researchers in a particular area might target clusters where their work and that of their target domain co-exist.

4.1 Subject Categories as Document Labels

We begin by estimating the extent to which conflated Web of Knowledge Subject Categories accurately describe the content of astrobiology publications. In Figure 4 we visualize the results of clustering the abstract data before sampling as described in Section 3.3. The same data is presented numerically in Table 5 in the online supplement. If SCs accurately reflect shared topical content of documents assigned to them, when the abstracts are clustered we should expect each SC to be primarily assigned a single cluster. However, when abstracts are assigned one of five clusters (Figure 4-top panel), we observe that the cluster membership for most SCs is heterogeneous: there is no clear correspondence between a cluster and a single dominant SC. Even the most common SC, Astronomy & Astrophysics, is primarily distributed across the first three clusters, but is represented in all five.

Table 5 does suggest some areas in which SCs may be more appropriate document labels. For example, Oceanography appears in only one cluster, and the Multidisciplinary Sciences SC is fairly evenly distributed across four of the five. However, when increasing the number of clusters to 10, 15, and 20 (Figure 4), the heterogeneity of SCs within an individual cluster becomes even more pronounced.

The dominance of the Astronomy & Astrophysics SC in the *conflated_SC_default* dataset suggests that we also examine the cluster relationships after the dataset has been sampled. Figure 5 shows the distribution of SCs in 5 clusters over 3 trials, where the cluster results of each trial are not related to each other. For example, in successive trials, the same abstract may be assigned to different clusters; for the Astrophysics & Geophysics SC, each trial results in different cluster assignments, though the overall distribution of clusters is roughly equal, suggesting that there is little variability between trials.

We present the data in Figure 5 as a fraction of the total number of features in the *conflated_SC_sampled* dataset, where the number of features representing each SC are equal (Figure 2). Furthermore, the data in Table 6 in the online supplement displays the data somewhat differently, where the fraction of features in each SC that are found in a given cluster is presented. Examining the data with the tables provided in the online supplement makes interpreting the results easier in some instances. In Figure 5 and Table 6, the Astronomy, Biochemistry & Microbiology, and Physics

14

SCs consistently cluster with their Multidisciplinary counterparts. Therefore, on the 5 cluster level, SCs seem to reasonably classify individual publications.

Since five clusters may not be sufficient to reflect the diversity of content within astrobiology, we increase the number of clusters in subsequent trials. One would intuitively expect more SC heterogeneity within each cluster; however, increasing the number of clusters also allows more potential of each SC to dominate a single cluster. When we increase the number of clusters to 10 (Figure 6, Table 7), we find that most of the SCs disperse into multiple clusters. One way to interpret this result is that more clusters allow finer distinctions between content to be revealed. For example, Physics and Physics-Multidisciplinary, which cluster together in each trial at the 5 cluster level, tend to cluster separately at the 10 cluster level. However, the Biochemistry & Microbiology SC and its multidisciplinary variant continue to have their abstracts cluster together. Moreover, from Figure 6, we observe that the abstracts in the Biotechnology & Applied Microbiology-Multidisciplinary SC consistently cluster together. At the 10 cluster level, more clusters contain single dominant SCs than at the 5 cluster level.

Figure 7 (Table 8) and Figure 8 (Table 9) present the results of clustering the abstracts into 15 and 20 clusters, respectively. We observe that many of the SCs are found distributed in multiple clusters. For example, at the 20 cluster level, what had been homogeneous cluster membership in the Biochemistry SCs at the 10 cluster level is split into three or more clusters, neither of which is shared across any other SC. Therefore, at these clustering levels, we operationalize a dominant SC within a cluster as one that either constitutes 50% or more of the abstracts alone, or one that is within 50% of the size of the most common SC[5]. By this approximation, the results at the 10 cluster level hold: as a group, the Biochemistry and Biotechnology-related SCs dominate the fewest clusters; the Astronomy, Oceanography and Physics group slightly more, and the Geochemistry and Geophysics SCs are again the most diverse, short of the Multidisciplinary Sciences SC. Overall, at the 10 cluster level, more clusters contain single dominant SCs than at the 5, 15 or 20 cluster levels, and the usefulness of SCs as document labels reaches a relative maximum.

In some cases, the trial processes reveal some inconsistencies in the cluster membership of SCs. For example, in the Biotechnology & Applied Microbiology-Multidisciplinary SC, one would

---

[5]For example, a cluster with SCs constituting 30%, 18%, 16% and 12% of the abstracts would have three dominant SCs.

expect to have diverse membership at the 15 cluster level (Figure 7). However, the Biotechnology & Applied Microbiology-Multidisciplinary SC is dominant in one cluster in trials 1 and 3, and is dominant in three clusters in trial 2. While these results may be an artifact of the sampling and multiple-trials processes, we would expect and find that the two related SCs, Biochemistry & Molecular Biology and Biochemistry & Molecular Biology-Multidisciplinary are found mostly within the same clusters. This observation also holds for the Geochemistry & Geophysics and Geochemistry & Geophysics-Multidisciplinary SCs. The multidisciplinary SC variants (BioChem & MBio, BioChem & MBio-M and GeoChem & GeoPhys, GeoChem & GeoPhys-M) are slightly more diverse than their associated core SC, but there is a high degree of similarity between the abstracts in these two sets of related SCs. Therefore, we conclude that even with some observed inconsistencies, the clusterer is collocating related abstracts across related SCs.

Certain related SCs tend to consistently cluster together, which suggests that SCs are sufficient for characterizing astrobiology publications. However, other SCs have a limited effectiveness as document labels in this interdisciplinary domain, as some SCs did not map well to successively smaller cluster sizes. Therefore, our results suggest that WoK SCs may not consistently reflect the diverse content of astrobiology publications.

4.2 Utilizing Subject Categories to Assess Interdisciplinarity

In this section, we attempt to leverage the heterogeneity of SC cluster membership to assess the interdisciplinarity of astrobiology publications, and analyze only the sampled data to de-emphasize the dominance of the Astronomy & Astrophysics SC in the dataset. Furthermore, many observations are similar to those discussed in Section 4.1; therefore, we will only mention in brief possible interdisciplinary connections that can be found by utilizing SCs.

In Figure 5, the clustering technique correctly collocates obviously similar SCs across all trials, but also identifies some less obvious potential interdisciplinary connections. For example, across all trials in Table 6, 9% of the abstract data in Biochemistry & Molecular Biology–Multidisciplinary clusters with 11-14% of the abstracts in Geochemistry & Geophysics. Given a document corpus of relatively equal SC distribution, as we have approximated here by the sampling process, these results suggest that papers from different SCs that consistently cluster together should be targeted for investigation by researchers interested in potential connections between the two fields.

Across all three trials at the 10 cluster level in Figure 6, a single clearly dominant SC could be identified in 27 of the 30 clusters. The Astronomy, Oceanography and Physics SCs demonstrated somewhat less monodisciplinary dominance at the 10 cluster level; all had roughly 20% of their abstracts assigned to other clusters. The Geochemistry & Geophysics and Environmental Sciences SCs demonstrated the most diversity apart from the pure Multidisciplinary Sciences SC, though somewhat surprisingly, the Geochemistry & Geophysics-Multidisciplinary SC appeared in fewer clusters than its core SC.

These findings yield several possible interpretations and applications. We would expect all astrobiology researchers to publish and cite primarily within their home disciplines, but as these results suggest, the norms of disciplinary diversity vary by field. A potential application of this approach is a field-specific baseline metric of interdisciplinarity, a method by which an individual's research output can be compared to others in the same field in terms of the potential interdisciplinary applicability of their work. This process could also result in an aggregate metric of interdisciplinarity for research teams via their past published work, while addressing the primary goal of discovering latent connections between the work of diverse researchers for the present and future.

Analyzing the heterogeneous cluster membership of publications from diverse SCs is one way to assess interdisciplinary research possibilities, but the probabilistic nature of this method should be emphasized. A heterogeneous cluster could indicate that SCs are poor document labels, or that the clustering level should be adjusted to better match the data and metadata, or that a potential interdisciplinary relationship exists. In either case, this process could inform targeted, iterative investigation.

4.3   Text Mining the Aggregated Abstracts

The sIB technique was employed to cluster the abstracts of the publications by the UHNAI team and the references within these publications. Figure 9, shows the results of clustering the data into 5 clusters. The results indicate that authors from their respective home disciplines cluster together (see Table 3 for the list of authors and their respective home disciplines). For example, the geologists Krot, Keil, Huss, Scott, and Jogo are strongly represented in cluster 4. One exception is Taylor (geologist) who clusters with the oceanographers (Cowen and Mottl). Additionally, Schörghofer (an astronomer by departmental affiliation) also clusters with the oceanographers. Furthermore, the

astrochemists (Bennett and Kaiser) have all of their publications in cluster 1. This result suggests that the sIB technique is able to cluster similar research on a high-level; however, utilizing more clusters should provide a lower-level view of overlap in research interests between the authors.

When running the sIB technique for 10 clusters, we begin to see where researchers may find potential collaboration opportunities, and we observe which authors have specialized or broad research interests. Research can be specialized but still integrate methods, techniques and data from multiple disciplines. We believe that an author who is represented primarily in a single cluster may not be engaging frequently in interdisciplinary research, or may be focusing on narrow research problems, or using similar research methods or equipment. In Figure 10, we see that the two astrochemists (Bennett and Kaiser) are entirely represented by cluster 8, consistent with the results presented in Figure 9. We know that their research is heavily influenced by their experimental apparati, thus suggesting that the experimental methods and apparati significantly affect the description of a research track. Interestingly, Schörghofer's research is on various planetary bodies such as Mars and the Moon, which is also true of Taylor. Therefore, clustering the text of the aggregated abstracts sufficiently illuminates similarities in research tracks across disciplinary boundaries, in this case, between astronomy and geology.

In Figure 11, we observe that Huss, Jewitt, Krot and Meech's research is found in many clusters. This signifies that their research is likely to be very interdisciplinary. With regards to those authors represented by a few clusters, we cannot conclude that their research is absolutely mono-disciplinary, as it may be very specialized, or utilize the same methods or apparati. However, we believe that those UHNAI authors with publications in multiple clusters are *more likely* to be engaged in interdisciplinary research. In Figure 12, we observe that of the senior (non-postdoctoral fellows) astronomers (Reipurth, Meech, Jewitt, Haghighipour, Owen, Schörghofer) half (Meech, Jewitt, and Owen) are fairly diverse in their research interests and the other half (Reipurth, Haghighipour, Schörghofer) are engaged in specialized or mono-disciplinary research.

These results suggest that the sIB method, in combination with aggregated abstracts, can illuminate areas of implicit commonality where the research areas of scientists from diverse disciplines overlap. Furthermore, while clusters do not inherently relate any information about a researcher's discipline, it is clear that researchers from the same department often cluster together. Therefore, we expect that performing a similar analysis on the entire NASA Astrobiology Institute will show

18

where collaborations between researchers can occur, and can assist NASA with outlining research priorities. These results can serve as the framework for a geospatial visualization of common yet unconnected research tracks and potential collaborators, similar to the "hot regions" described by Bornmann and Waltman (2011).

5   Discussion and Conclusions

We clustered astrobiology abstract data to evaluate SCs as document labels. We attempt to reconcile clustering (bottom-up approach) with pre-defined categories (top-down approach). The clusters produced by text mining the abstract data did not generally correspond well to the SCs. Therefore, we conclude that SCs are not well suited to the classification of astrobiology publications, and speculate that this may also be true for other interdisciplinary fields. One explanation is that astrobiology research outputs cite mono-disciplinary and interdisciplinary publications which may prevent SCs from forming cohesive clusters. Additionally, as discussed in Small (2010), many journals publish highly diverse content, which no journal-level classification system could represent completely. The class imbalance problem in our dataset requires us to explore utilizing an oversampling technique. While we believe that the method remedies the skewed distribution of conflated SCs in our dataset, performing a text mining clustering analysis on a balanced astrobiology dataset without oversampling may produce different results. That is, SCs may be more accurate when the distribution of SCs is uniform without oversampling using synthetic data. Nonetheless, the distribution of departmental affiliations of the UHNAI researchers is skewed, which affects the distribution of publications across different SCs; it is likely that this scenario will be consistent with the other NASA Astrobiology Institute teams.

Our results suggest that 10 clusters may be the most appropriate level at which to analyze the astrobiology collection (Figure 6). Too few clusters and the interdisciplinary diversity of the source documents is not well represented; too many and they may be oversegregated, lessening the chance to identify potential commonalities in documents from different disciplines and SCs. We suggest that when documents from different SCs cluster together, this may indicate implicit interdisciplinary connection, where knowledge in one field might inform another. Having researchers from the constituent disciplines evaluate these common documents may provide one mechanism by

which interdisciplinary science can take place, and provide a starting point for potentially productive interdisciplinary collaborations.

Similarly, text mining the aggregated abstracts using the sIB method is also suited to the task of finding collaboration opportunities. Our experiments consistently showed that authors from the same academic department tended to have their publications cluster together. If this were not the case, we would be unable to make any claims regarding the similarity of publications within a given cluster. We suggest that authors whose publications cluster together could collaborate productively. An author that has publications in many clusters indicates that they are engaged in interdisciplinary research, or perhaps that they are not, but should be. Those authors with few publications were either underrepresented in WoK and ADS, or were post-doctoral fellows at the UHNAI. We find that the majority of publications by UHNAI investigators and post-doctoral fellows appear in multiple clusters, providing evidence of actual or potential interdisciplinary research. This is an encouraging result, as promoting boundary-crossing scientific research is one of the goals of the NASA Astrobiology Institute. Younger generations of researchers will need to synthesize techniques from multiple disciplines to answer some of the most fundamental questions in science in general, and astrobiology in particular.

We insinuated that a strong conclusion cannot be made regarding those authors that are strongly represented in a single cluster. Research in this context is either: 1) interdisciplinary but specialized, perhaps incorporating a synthesis between methods, techniques and data from multiple disciplines, but with a narrow scope or 2) mono-disciplinary. Distinguishing between these two cases requires studying the individual works in each cluster. Additionally, such an analysis would lead to narrowing the scope of collaboration between two or more researchers that are found within a single cluster. This analysis will be explored in future work.

The context of the interdisciplinary field of astrobiology has permitted us to explore a method of measuring interdisciplinarity, and identify potential collaboration opportunities. We find that most of the UHNAI team are engaged in interdisciplinary research, and that our method suggests where productive interdisciplinary collaborations could occur. We believe our method, which combines bibliometrics and machine learning, makes valid predictions, based on our a priori knowledge of the structure of the research team and those intra-team collaborations that currently exist. Bibliometric studies of interdisciplinarity can benefit when augmented with machine learning algorithms, in an

attempt to understand the fine-grained details of interdisciplinary research.

Acknowledgements

References

Bjurström, A., & Polk, M. (2011). Climate change and interdisciplinarity: a co-citation analysis of IPCC Third Assessment Report. *Scientometrics*, 87, 525–550.

Bornmann, L., & Waltman, L. (2011). The detection of "hot regions" in the geography of science–A visualization approach by using density maps. *Journal of Informetrics*, 5(4), 547–553.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64, 351–374.

Brewer, G. D. (1999). The challenges of interdisciplinarity. *Policy Sciences*, 32, 327–337.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357.

Cockell, C. (2002). Astrobiology–a new opportunity for interdisciplinary thinking. *Space Policy*, 18(4), 263–266.

Derrick, G., Sturk, H., Haynes, A., Chapman, S., & Hall, W. (2010). A cautionary bibliometric tale of two cities. *Scientometrics*, 84, 317–320.

Gargaud, M., & Tirard, S. (2011). Exobiology: An Example of Interdisciplinarity at Work. In Jean-Pierre Lasota, editor, *Astronomy at the Frontiers of Science*, volume 1 of *Integrated Science & Technology Program*, pages 337–350. Springer Netherlands.

Jacsó, P. (2005). As we may search: comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*, 89, 1537–1547.

Kostoff, R. (1998). The use and misuse of citation analysis in research evaluation. *Scientometrics*, 43, 27–43.

Kostoff, R., del Río, J. A., Humenik, J. A., García, E. O., & Ramírez, A. M. (2001). Citation mining: Integrating text mining and bibliometrics for research user profiling. *Journal of the American Society for Information Science and Technology*, 52(13), 1148–1156.

Kousha, K., & Thelwall, M. (2008). Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics*, 74, 273–294.

Morillo, F., Bordons, M., & Gómez, I. (2001). An approach to interdisciplinarity through bibliometric indicators. *Scientometrics*, 51, 203–222.

National Academies. Committee on Facilitating Interdisciplinary Research, of the Committee on Science, Engineering, and Public Policy. (2005). *Facilitating Interdisciplinary Research*. Washington, DC: National Academies Press.

National Science Foundation. Introduction to interdisciplinary research. `http://www.nsf.gov/od/oia/additional_resources/interdisciplinary_research/`, Accessed November 21, 2011.

Oliver, C. A., & Fergusson, J. (2007). Astrobiology: A pathway to adult science literacy? *Acta Astronautica*, 61(7-8), 716–723.

Porter, A., Cohen, A., Roessner, J. D., & Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics*, 72, 117–147.

Porter, A., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, 81, 719–745.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130–137.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 82, 263–287.

Slonim, N., Friedman, N., & Tishby, N. (2002). Unsupervised document classification using sequential information maximization. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 129–136, New York, NY, USA.

Small, H. (2010). Maps of science as interdisciplinary discourse: co-citation contexts and the role of analogy. *Scientometrics*, 83, 835–849.

Staley, J. (2003). Astrobiology, the transcendent science: the promise of astrobiology as an integrative approach for science and engineering education and research. *Current Opinion in Biotechnology*, 14(3), 347–354.

Sugimoto, C. (2011). Looking across communicative genres: a call for inclusive indicators of interdisciplinarity. *Scientometrics*, 86, 449–461.

Upham, S., & Small, H. (2010). Emerging research fronts in science and technology: patterns of new knowledge development. *Scientometrics*, 83, 15–38.

van Leeuwen, T. N. (2007). Modelling of bibliometric approaches and importance of output verification in research performance assessment. *Research Evaluation*, 16(2), 93–105.

van Raan, A. F. J., & van Leeuwen, T. N. (2002). Assessment of the scientific basis of interdisciplinary, applied research: Application of bibliometric methods in Nutrition and Food Research. *Research Policy*, 31(4), 611–632.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., Rafols, I., & Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, 5(1), 14–26.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques, 2nd edition.* Morgan Kaufmann, San Francisco.

Zhang, J., Vogeley, M. S., & Chen, C. (2011). Scientometrics of big science: a case study of research in the Sloan Digital Sky Survey. *Scientometrics*, 86, 1–14.

Zhang, L., Liu, X., Janssens, F., Liang, L., & Glänzel, W. (2010). Subject clustering analysis based on ISI category classification. *Journal of Informetrics*, 4(2), 185–193.

Attribute/Word

Abstract	Class/SC	water	is	found	on	the	earth	and	in	solar	system	exists	moon	mars
1	Astro	1	1	1	1	2	1	1	1	1	1	0	0	0
2	Astro	1	0	0	1	1	0	1	0	0	0	1	1	1

Figure 1: Depiction of feature vectors as constructed from abstract data.
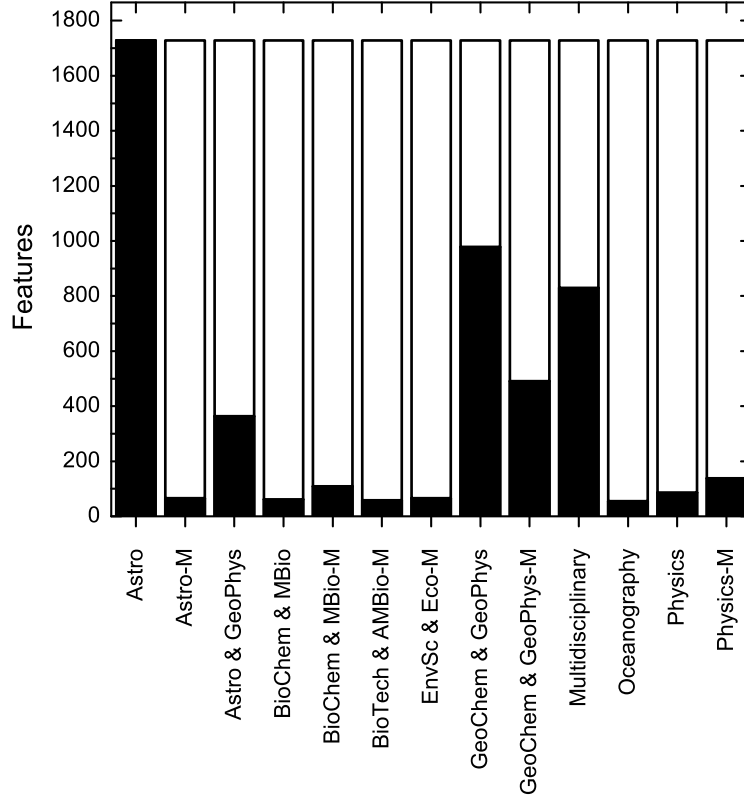
Figure 2: The distribution of real (black) and synthetic (white) data in the *conflated_SC_sampled* dataset.
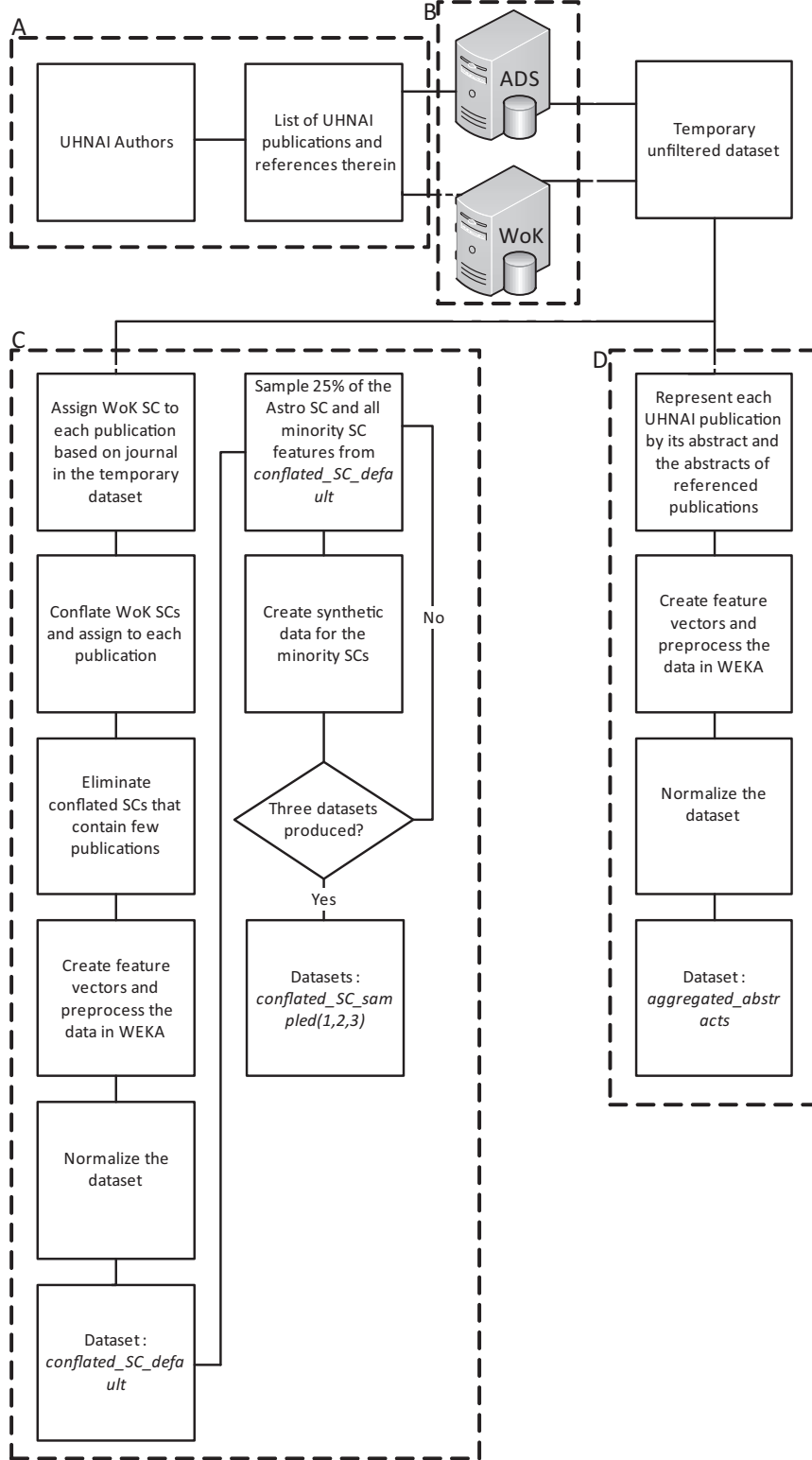
Figure 3: An overview of the steps in our methodology. The UHNAI authors and publications to be harvested are are shown as region A and outlined in Section 3. The motivation for using the sIB is discussed in Section 3.1. The data collection procedure is discussed in Section 3.2 and is described by region B. The method for creating our conflated subject categories is shown as region C and discussed in Sections 3.3 and 3.5. The method for creating our aggregated abstracts is shown as region D and discussed in Sections 3.4 and 3.5.
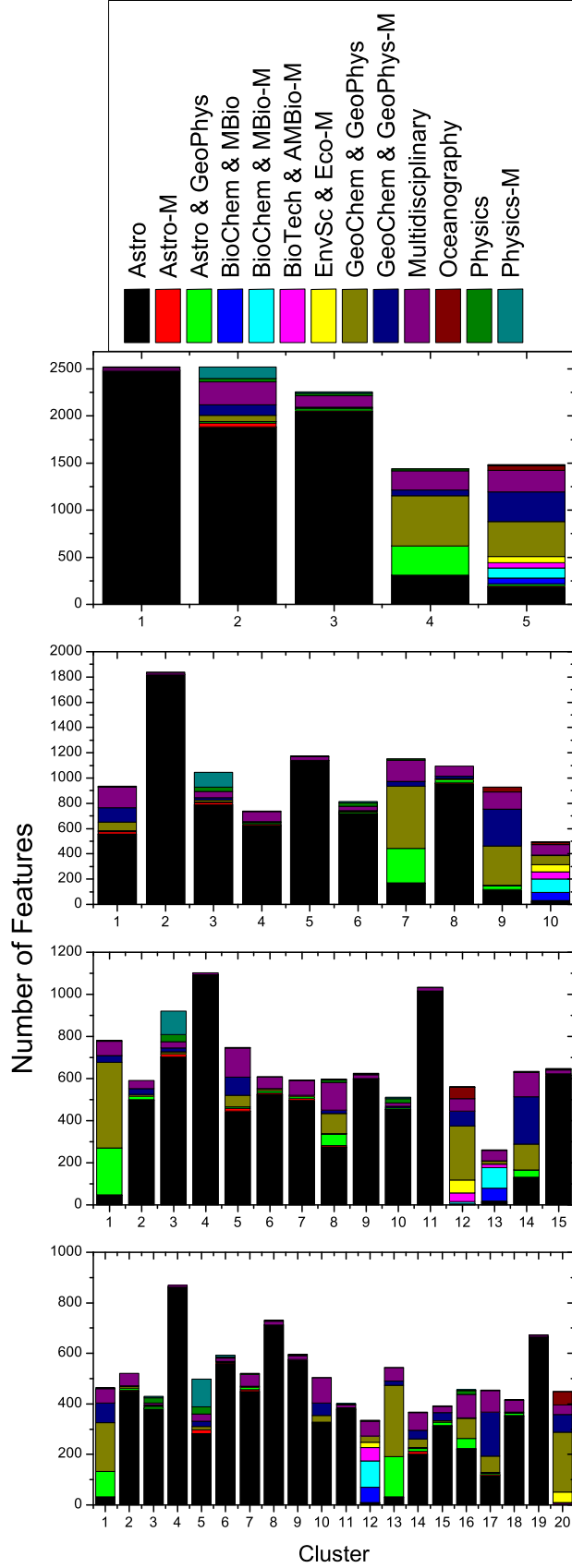
Figure 4: The results of clustering the *conflated_SC_default* dataset. Results are given for 5, 10, 15 and 20 clusters from top to bottom.
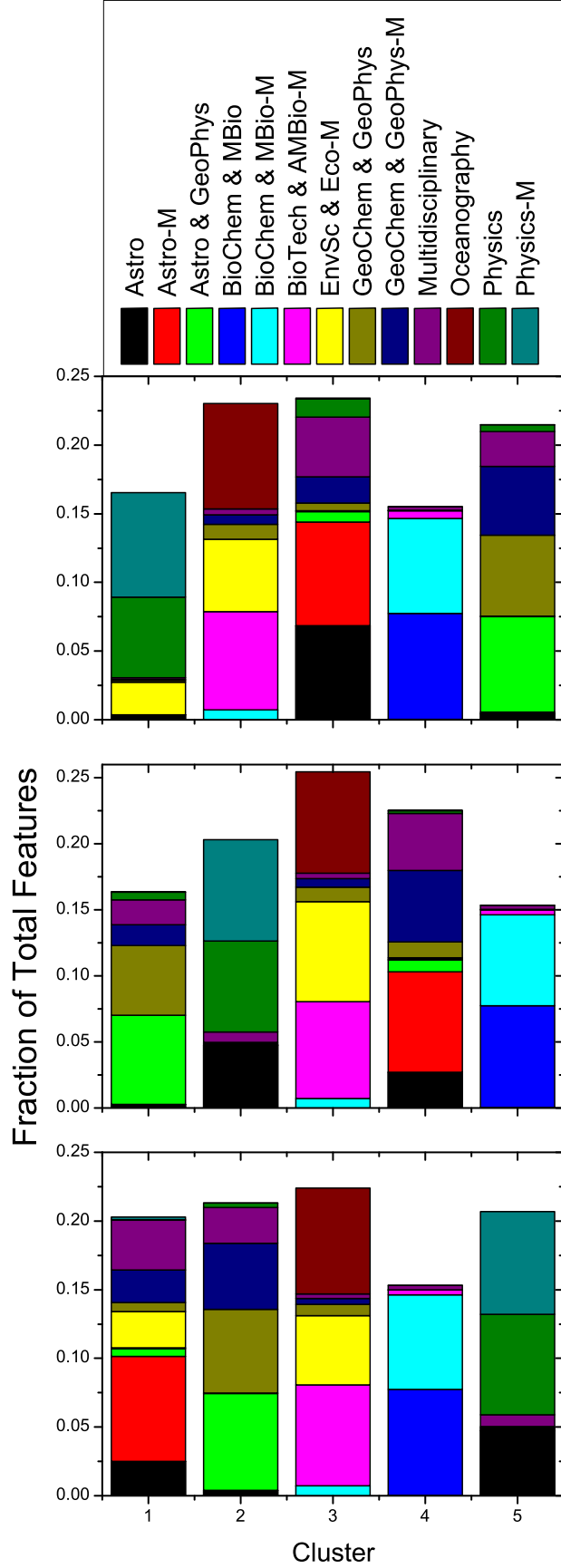
Figure 5: The results of clustering the *conflated_SC_sampled* dataset in three separate trials. Each abstract is assigned one of 5 clusters.
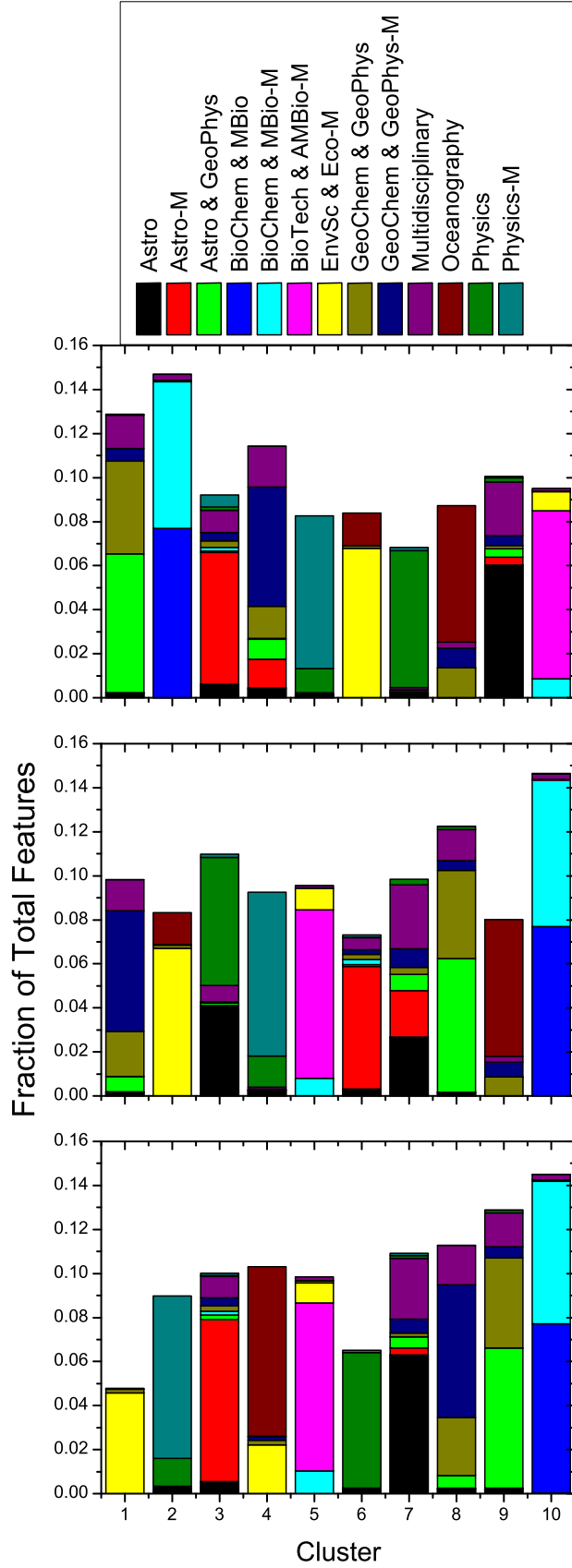
Figure 6: The results of clustering the *conflated_SC_sampled* dataset in three separate trials. Each abstract is assigned one of 10 clusters.
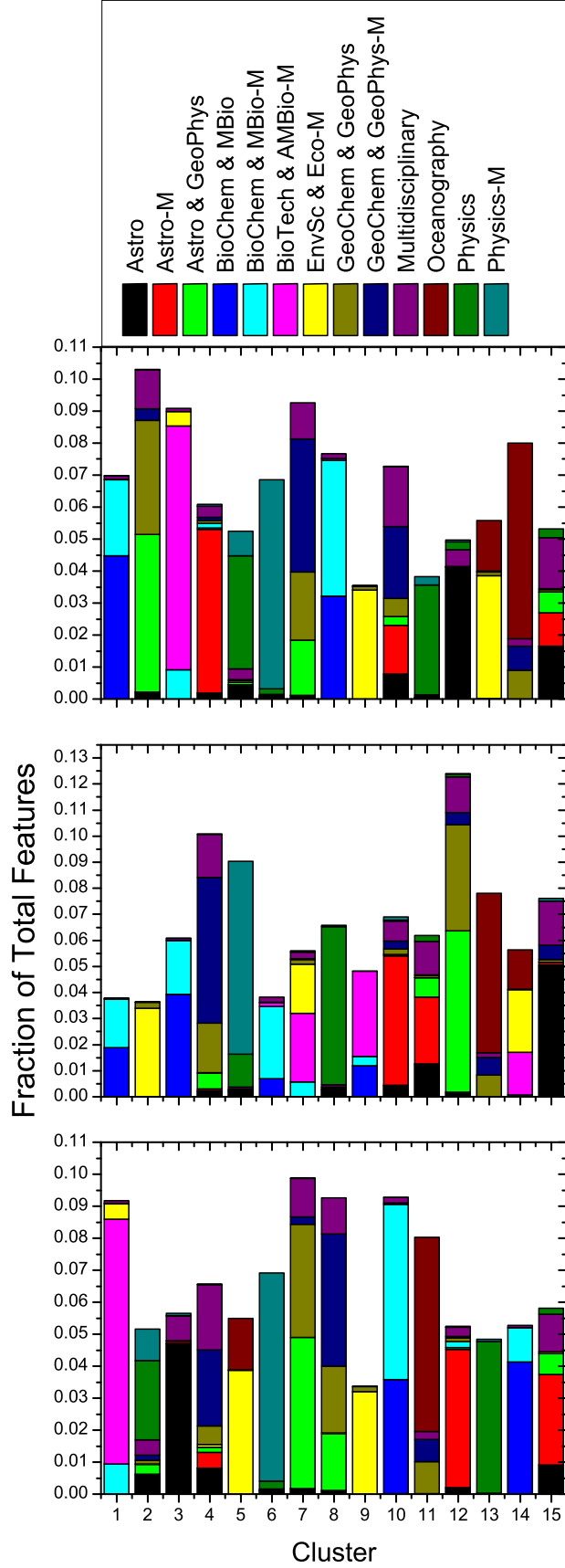
Figure 7: The results of clustering the *conflated_30C_sampled* dataset in three separate trials. Each abstract is assigned one of 15 clusters.
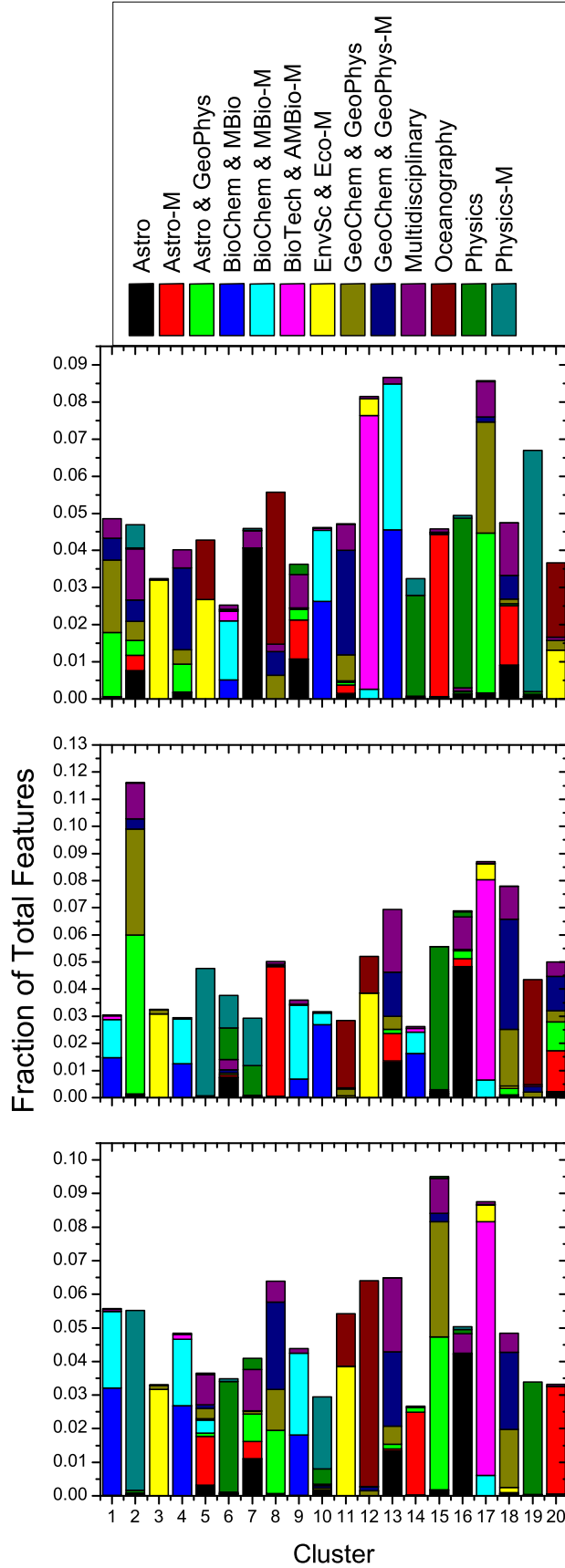
Figure 8: The results of clustering the *conflated_31C_sampled* dataset in three separate trials. Each abstract is assigned one of 20 clusters.
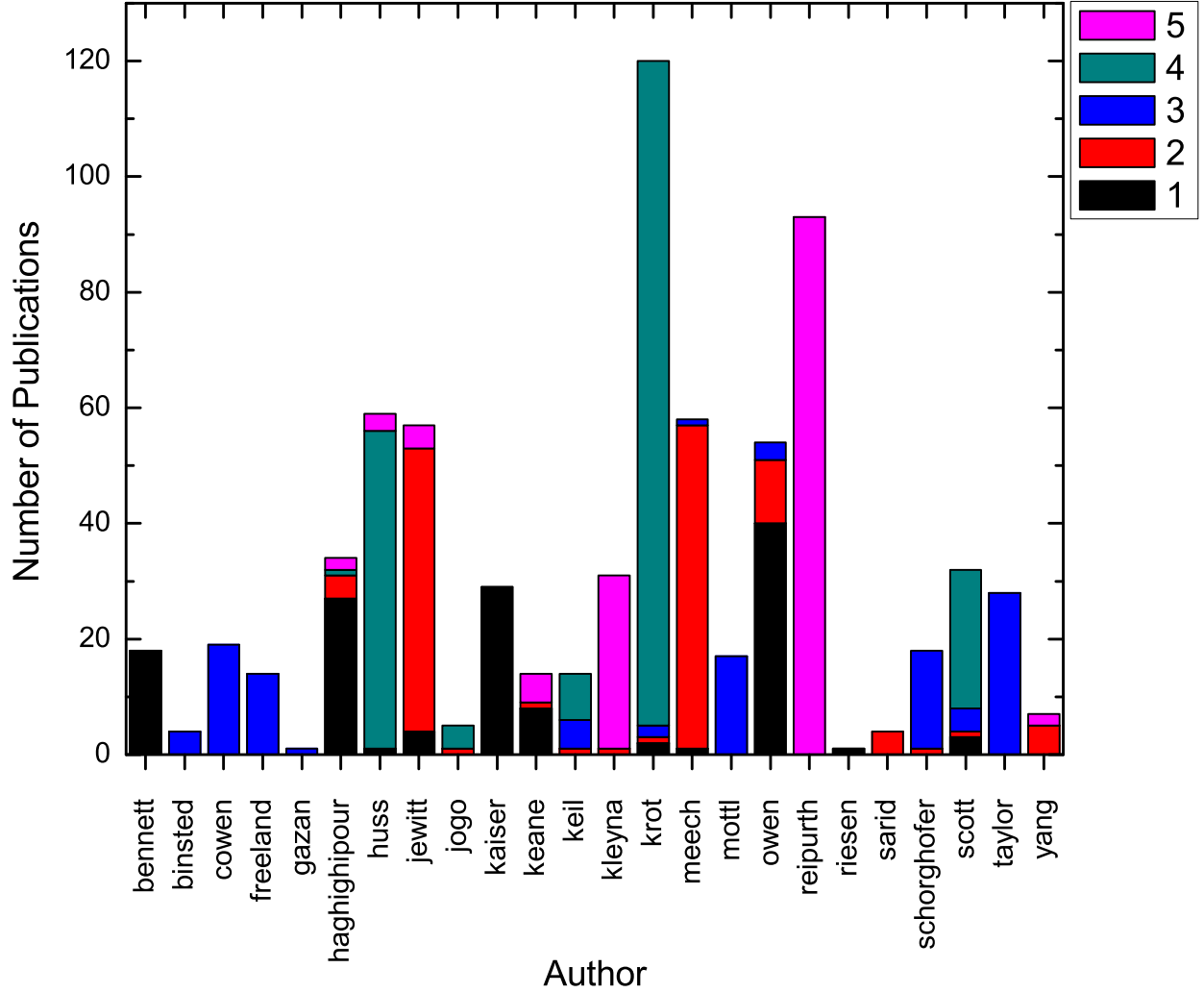
Figure 9: Clustering the *aggregated_abstracts* dataset using 5 clusters. This plot ensures that we are not obtaining extremely unlikely correlations and shows that researchers from the same academic department are largely clustering together. For example, Bennett is a post-doctoral fellow working with Kaiser; they often publish together and their aggregated abstracts are clustering entirely in cluster 1. As another example, the geologists/geophysists Krot, Keil, Huss, Scott and Jogo are all strongly represented in cluster 4. The one exception is Taylor, who appears to be clustering more strongly with the two oceanographers (Cowen and Mottl). As expected, researchers have the most in common with those in their home discipline.
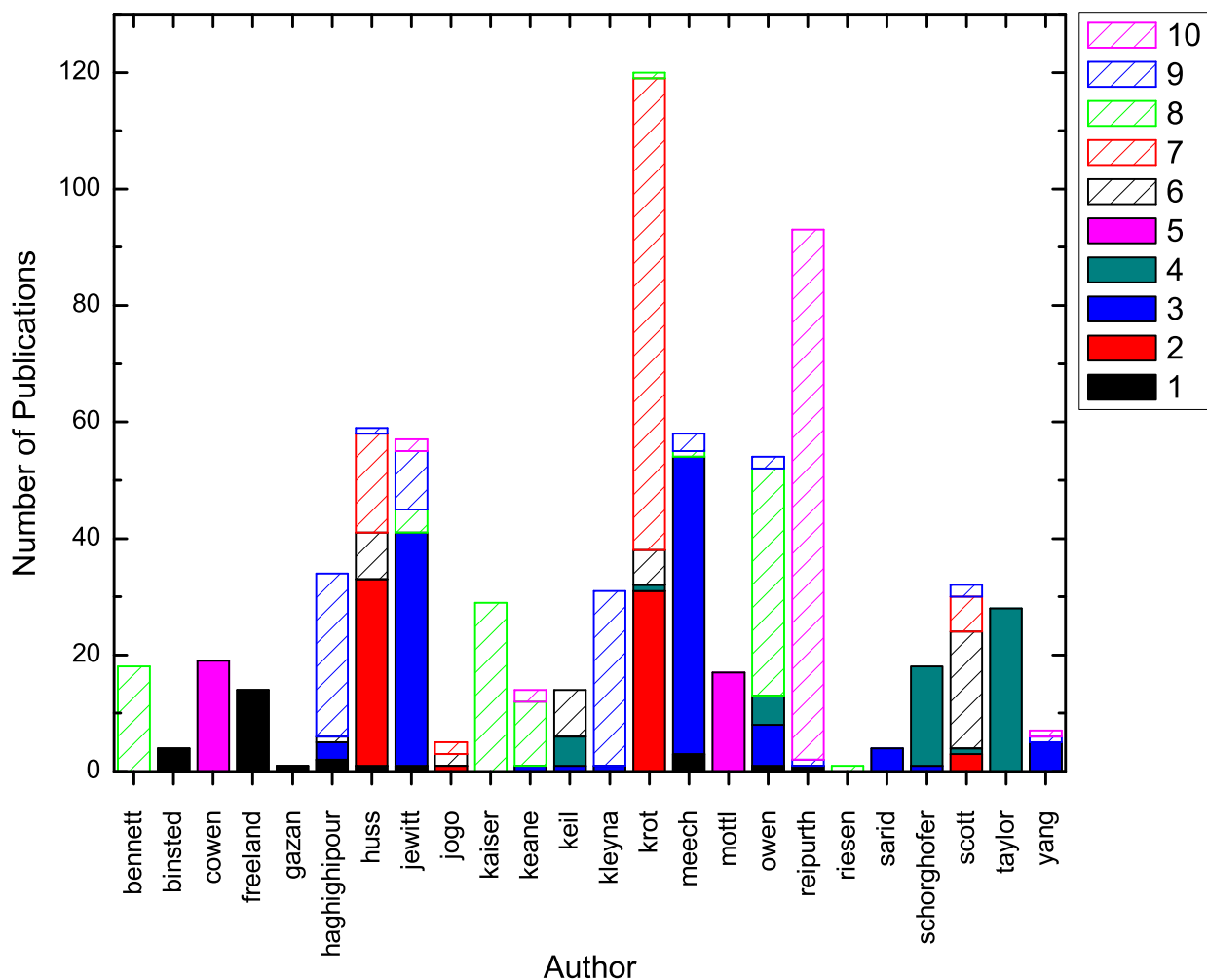
Figure 10: Clustering the *aggregated_abstracts* dataset using 10 clusters. The two oceanographers (Cowen and Mottl) have all of their papers clustering together in Cluster 5. The same is true of Bennett and Kaiser (Astrochemistry). In the previous figure, Taylor was clustering with the oceanographers. However, we can see here that Taylor's work is similar to that of Schörghofer's, despite their membership to different home disciplines (Geology, and Astronomy respectively). Rather striking is the mono-disciplinarity regarding Reipurth's research.
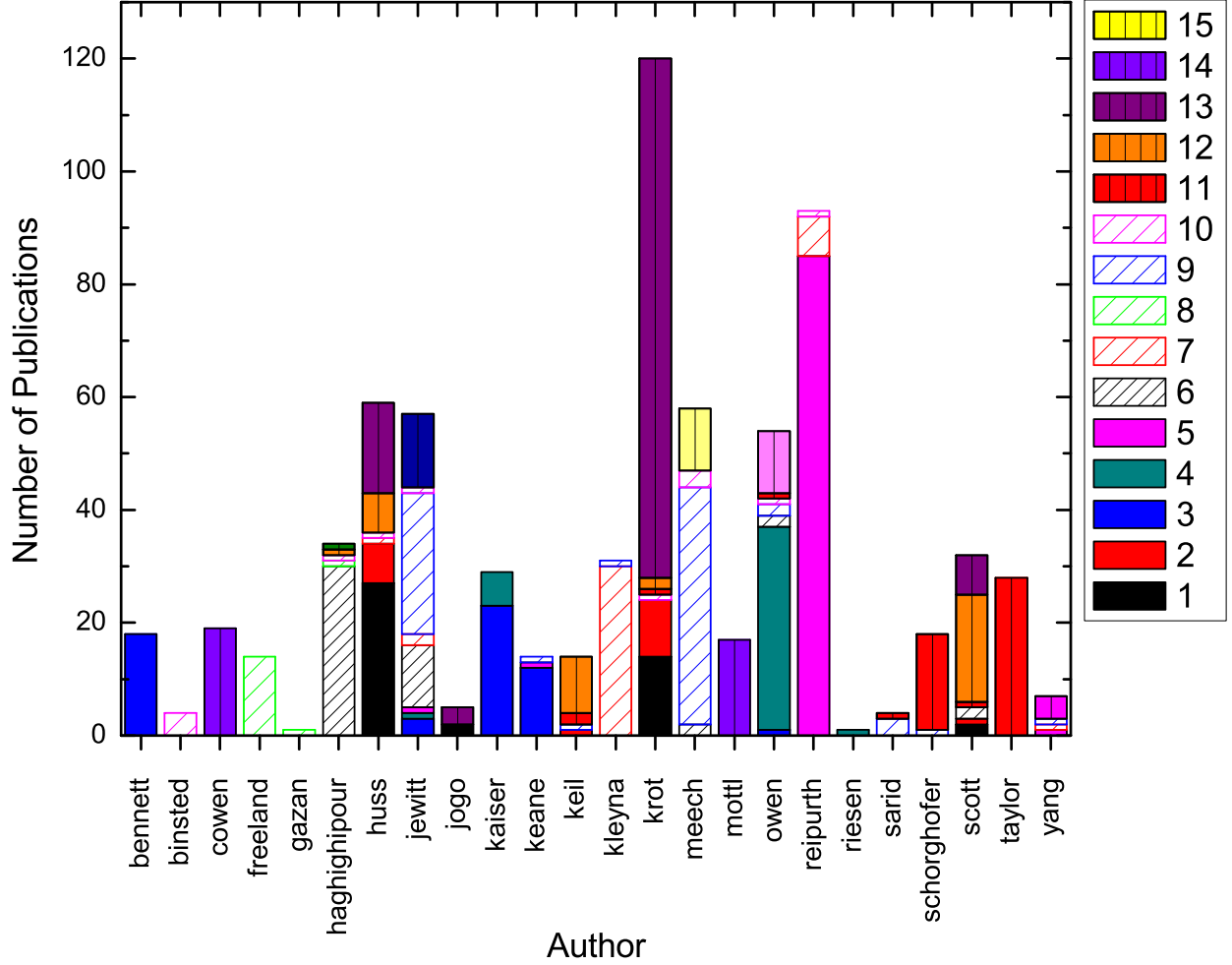
Figure 11: Clustering the *aggregated_abstracts* dataset using 15 clusters. In this figure, Bennett and Kaiser are no longer entirely represented by a single cluster. When we utilized 5 and 10 clusters, Binsted and Gazan (Computer Science) and Freeland (Biology) had their publications cluster together. We know in particular that the research by the computer scientists is likely to be the most dissimilar to all authors from other home disciplines. However, when clustering with 15 clusters, we observe Binsted's research depart from the cluster that contains Gazan and Freeland's research and that the research has a tangential relation to research produced by other team members.
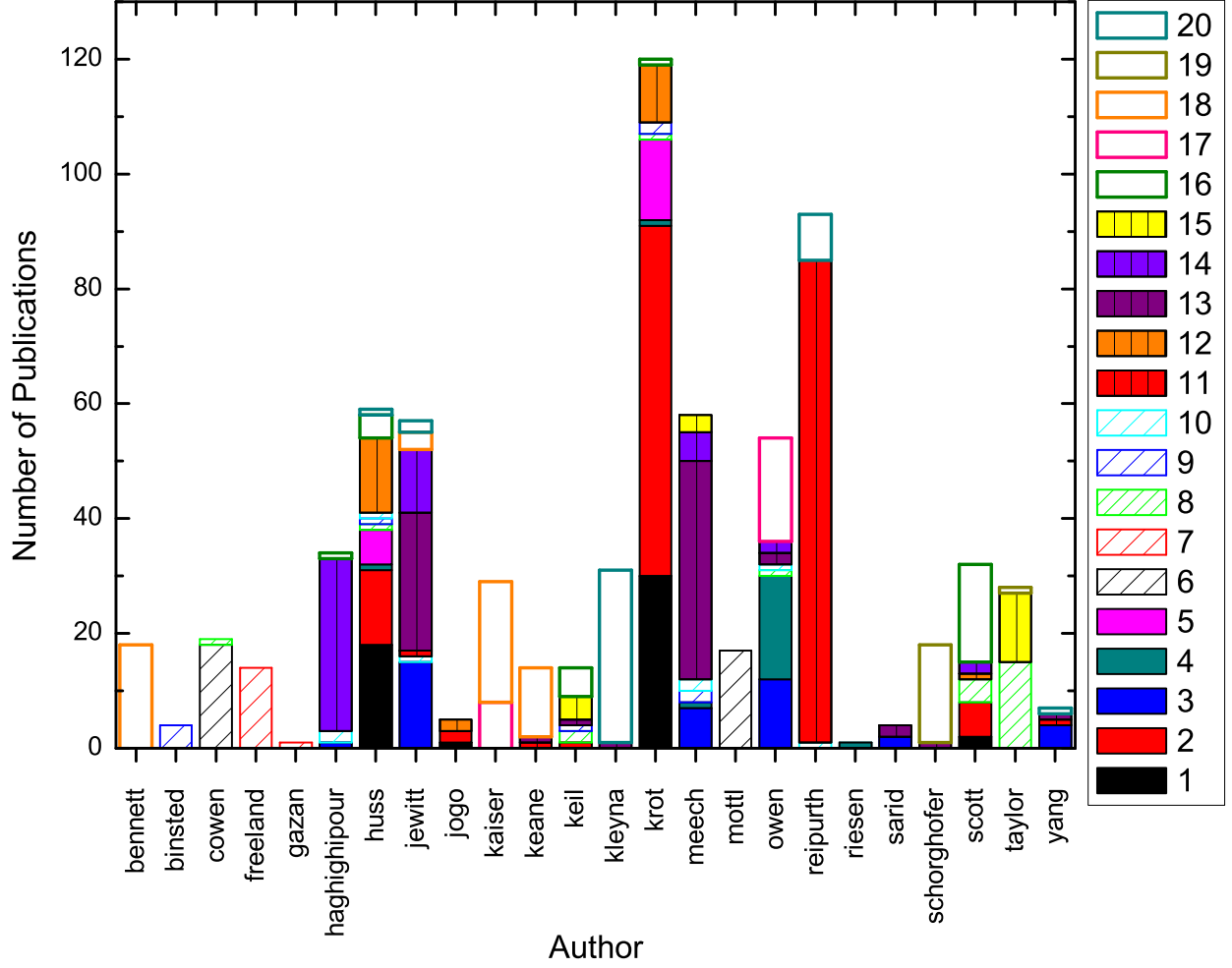
Figure 12: Clustering the *aggregated_abstracts* dataset using 20 clusters. If we assume that membership to many clusters indicates a high degree of interdisciplinarity, Huss is the most interdisciplinary UHNAI team member. Of the senior Astronomers (Reipurth, Meech, Jewitt, Haghighipour, Owen, Schörghofer) half (Meech, Jewitt, and Owen) are fairly diverse in their research interests, or engage in IDR, and the other half (Reipurth, Haghighipour, Schörghofer) are engaged in specialized or mono-disciplinary research. As younger researchers, the UHNAI post-doctoral fellows appear to be engaging in interdisciplinary research.

Table 1: Mapping of WoK subject categories to conflated subject categories.

Conflated Subject Category	WoK Source Subject Category
Agricultural Engineering-Multidisciplinary	Agricultural Engineering **with** Fisheries
Astronomy & Astrophysics	Astronomy & Astrophysics
Astronomy & Astrophysics-Multidisciplinary	Astronomy & Astrophysics-Multidisciplinary **or** Astronomy & Astrophysics **with**: Mechanics; Engineering-Aerospace; History; Multidisciplinary Sciences; Physics; Spectroscopy
Astrophysics & Geophysics	Astronomy & Astrophysics **with**: Geochemistry & Geophysics; Geosciences-Multidisciplinary
Biochemistry & Molecular Biology	Biochemistry & Molecular Biology
Biochemistry & Molecular Biology-Multidisciplinary	Biochemistry & Molecular Biology **with**: Biochemical Research Methods; Chemistry, Analytical; Biotechnology & Applied Microbiology; Mathematical & Computational Biology; Biology; Biophysics; Cell Biology; Computer Science, Interdisciplinary Applications; Genetics & Heredity; Medicine, Research & Experimental; Chemistry, Analytical; Chemistry, Medicinal; Chemistry, Organic; Pharmacology & Pharmacy; Evolutionary Biology; Microbiology; Immunology; Infectious Diseases
Biology	Biology
Biology-Multidisciplinary	Biology **with**: Ecology; Evolutionary Biology; Environmental Sciences; Mathematical & Computational Biology
Biology & Geology	Biology **or** Environmental Sciences **with** Geosciences, Multidisciplinary
Biotechnology & Applied Microbiology	Biotechnology & Applied Microbiology
Biotechnology & Applied Microbiology-Multidisciplinary	Biotechnology & Applied Microbiology **with**: Food Science & Technology; Microbiology; Genetics & Heredity; Marine & Freshwater Biology
Chemistry	Chemistry; Chemistry, Analytical; Chemistry, Physical; Chemistry, Organic; Chemistry, Inorganic & Nuclear

Conflated Subject Category	WoK Source Subject Category
Chemistry-Multidisciplinary	Chemistry, Multidisciplinary **or** Chemistry **or** Chemistry, Analytical **with**: Spectroscopy; Chemistry, Medicinal; Nanoscience & Nanotechnology; Materials Science, Multidisciplinary; Chemistry, Applied; Computer Science, Information Systems; Computer Science, Interdisciplinary Applications; Pharmacology & Pharmacy; Environmental Sciences; Toxicology; Physics, Condensed Matter; Engineering, Chemical; Mathematics & Computational Biology; Oceanography; Nuclear Science & Technology; Polymer Science
Chemistry & Physics	Chemistry **with** Physics
Chemistry & Physics-Multidisciplinary	Chemistry and Physics **with** Nuclear Science & Technology
Computer Science	Computer Science
Computer Science-Multidisciplinary	Computer Science, Multidisciplinary **or** Computer Science **with**: Information Science & Library Science; Cybernetics; Computer Science, Artificial Intelligence; Computer Science, Theory & Methods; Engineering, Electrical & Electronic; Computer Science, Hardware & Architecture; Computer Science, Information Systems; Computer Science, Interdisciplinary Applications; Geosciences, Multidisciplinary; Physics, Fluids & Plasmas
Crystallography	Crystallography
Education	Education **or** Education **with** Multidisciplinary Sciences
Engineering	Engineering; Engineering, Instruments & Instrumentation; Engineering, Electrical & Electronic; Engineering, Mechanical
Environmental Sciences & Ecology	Environmental Sciences **or** Ecology
Environmental Sciences & Ecology-Multidisciplinary	Environmental Sciences **or** Ecology **with**: Limnology; Evolutionary Biology; Marine & Freshwater Biology; Microbiology; Oceanography; Engineering, Civil; Water Resources; Engineering, Environmental; Engineering, Chemical; Geology; Meteorology & Atmospheric Sciences; Geography, Physical; Geosciences, Multidisciplinary; Soil Science; Toxicology; Fisheries
Genetics & Heredity	Genetics & Heredity
Genetics & Heredity-Multidisciplinary	Genetics & Heredity **with**: Ecology; Evolutionary Biology
Geochemistry & Geophysics	Geochemistry & Geophysics

Conflated Subject Category	WoK Source Subject Category
Geochemistry & Geophysics-Multidisciplinary	Geochemistry & Geophysics-Multidisciplinary **or** Geochemistry & Geophysics **with**: Geology; Meteorology & Atmospheric Sciences; Mineralogy; Geography, Physical; Geosciences, Multidisciplinary; Paleontology
Geography	Geography, Physical
Geology	Geology
Geology-Multidisciplinary	Geology **or** Geosciences, Multidisciplinary **with**: Energy & Fuels; Engineering, Petroleum; Mineralogy; Mining & Mineral Processing; Paleontology; Geography, Physical; Mathematics, Interdisciplinary Applications
Geology & Oceanography	Geology **with** Oceanography
Geophysics & Oceanography	Geochemistry & Geophysics **with** Oceanography
Instruments & Instrumentation	Instruments & Instrumentation
Life Sciences & Biomedicine-Multidisciplinary	Life Sciences & Biomedicine, Other Topics; Multidisciplinary Sciences; Science & Technology, Other Topics
Materials Science	Materials Science
Materials Science-Multidisciplinary	Materials Science, Multidisciplinary **or** Materials Science **with** Physics, Metallurgy & Metallurgical Engineering
Mathematical & Computational Biology	Mathematical & Computational Biology
Mathematics	Mathematics; Mathematics, Applied; Statistics & Probability
Medicine	Medical Sciences; Psychology, Clinical; Medicine, General & Internal; Public, Environmental & Occupational Health; Sport Sciences
Meteorology & Atmospheric Sciences	Meteorology & Atmospheric Sciences
Meteorology & Oceanography	Meteorology & Atmospheric Sciences **with** Oceanography
Microbiology	Microbiology
Mineralogy	Mineralogy
Multidisciplinary Sciences	Multidisciplinary Sciences
Neurosciences	Neurosciences
Nutrition & Dietetics	Nutrition & Dietetics
Oceanography	Oceanography
Oceanography & Marine Biology	Oceanography **and** Marine & Freshwater Biology **or** Limnology
Optics	Optics
Optics-Multidisciplinary	Optics **with**: Spectroscopy; Engineering, Multidisciplinary
Pharmacology & Pharmacy	Pharmacology & Pharmacy

Conflated Subject Category	WoK Source Subject Category
Physics	Physics; Physics, Fluids & Plasmas
Physics-Multidisciplinary	Physics, Multidisciplinary **or** Physics **with**: Mechanics; Physics, Particles & Fields; Physics, Nuclear; Nuclear Science & Technology; Physics, Atomic, Molecular & Chemical; Chemistry, Physical; Instruments & Instrumentation; Optics; Thermodynamics; Energy & Fuels
Psychology	Psychology
Spectroscopy	Spectroscopy
Spectroscopy-Multidisciplinary	Spectroscopy **with**: Chemistry, Physical; Chemistry, Analytical; Physics, Atomic, Molecular & Chemical
Virology	Virology
Zoology	Zoology

Table 2: The distribution of conflated subject categories, their corresponding abstracts and the fraction of abstracts obtained through ADS and WoK.

Subject Category	Number of abstracts	Fraction of dataset (%)	ADS (%)	WoK (%)
Astronomy & Astrophysics [Astro]	6914	67.68	98.7	1.3
Astronomy & Astrophysics-Multidisciplinary [Astro-M]	66	0.65	98.5	1.5
Astrophysics & Geophysics [Astro & GeoPhys]	364	3.56	93.4	6.6
Biochemistry & Molecular Biology [BioChem & MBio]	61	0.6	0	100
Biochemistry & Molecular Biology-Multidisciplinary [BioChem & MBio-M]	109	1.07	0.9	99.1
Biotechnology & Applied Microbiology-Multidisciplinary [BioTech & AMBio-M]	58	0.57	0	100
Environmental Sciences & Ecology-Multidisciplinary [EnvSc & Eco-M]	66	0.65	1.5	98.5
Geochemistry & Geophysics [GeoChem & GeoPhys]	978	9.6	65.1	34.9
Geochemistry & Geophysics-Multidisciplinary [GeoChem & GeoPhys-M]	491	4.8	45.8	54.2
Multidisciplinary Sciences [Multidisciplinary]	830	8.12	78.9	21.1
Oceanography	55	0.54	0	100
Physics	86	0.84	100	0
Physics-Multidisciplinary [Physics-M]	138	1.35	98.6	1.4

Table 3: Home discipline of the authors at the University of Hawaii NASA Astrobiology Institute. An asterisk (*) denotes a post-doctoral researcher.

Author	Departmental Affiliation/Home Discipline
Bennett*	Chemistry
Binsted	Computer Science
Cowen	Oceanography
Freeland	Biology
Gazan	Computer Science
Haghighipour	Astronomy
Huss	Geology
Jewitt	Astronomy
Jogo*	Geology
Kaiser	Chemistry
Keane*	Astronomy
Keil	Geology
Kleyna*	Astronomy
Krot	Geology
Meech	Astronomy
Mottl	Oceanography
Owen	Astronomy
Reipurth	Astronomy
Riesen*	Astronomy
Sarid*	Astronomy
Schörghofer	Astronomy
Scott	Geology
Taylor	Geology
Yang*	Astronomy

Table 4: Statistics for the UHNAI aggregated abstracts dataset. A sample ($N = 100$) of the total number of UHNAI publications is shown to estimate the completeness of the aggregated abstracts.

	ADS	WoK	Total
Total UHNAI publications in the dataset	655 (89.6%)	76 (10.4%)	731
Number of publications randomly selected	88	12	100
Total number of references found across the sample	3426	756	4182
Total referenced abstracts harvested	2908	386	3294
Average completeness of the aggregated abstracts[1]	77.5%	50.3%	74.3%

[1]The average completeness is measured as the mean of the completeness of each individual aggregated abstract in the sample.

Online Supplement

Table 5: Clustering the *conflated_SC_default* dataset with 5, 10, 15, and 20 clusters.

	Astro	Astro-M	Astro & GeoPhys	BioChem & MBio	BioChem & MBio-M	BioTech & AMBio-M	EnvSc & Eco-M	GeoChem & GeoPhys	GeoChem & GeoPhys-M	Multidisc-iplinary	Oceano-graphy	Physics	Physics-M
5 Clusters													
1	2484	0	0	0	0	0	0	0	2	30	0	4	1
2	1887	38	17	0	1	0	0	66	112	244	0	34	120
3	2048	13	20	0	1	0	0	6	4	123	0	27	12
4	308	2	310	0	1	0	1	532	62	204	0	16	2
5	187	13	17	61	106	58	65	374	311	229	55	5	3
10 Clusters													
1	558	17	6	0	0	0	1	68	116	163	1	2	3
2	1821	0	0	0	0	0	0	0	0	15	0	0	1
3	788	17	6	0	0	0	0	14	19	50	0	34	116
4	626	10	12	0	0	0	1	3	2	79	0	6	0
5	1141	1	0	0	0	0	0	0	0	27	0	3	3
6	719	2	9	0	1	0	1	8	4	32	0	28	11
7	168	2	274	0	0	0	0	492	39	166	0	11	1
8	959	8	22	0	0	0	0	9	19	78	0	0	0
9	113	3	30	0	0	0	6	312	288	140	35	1	0
10	21	6	5	61	108	58	57	72	4	80	19	1	3
15 Clusters													
1	46	2	222	0	0	0	0	408	30	70	0	3	0
2	498	2	17	0	0	0	0	7	28	40	0	0	0
3	702	15	1	0	0	0	0	9	19	28	0	34	113
4	1093	0	0	0	0	0	0	0	0	9	0	0	0
5	445	15	5	0	0	0	0	56	85	139	0	1	1
6	525	6	10	0	0	0	0	9	3	53	0	2	0
7	497	9	11	0	0	0	0	3	1	70	0	3	0
8	276	5	56	0	1	0	1	94	17	131	0	12	5
9	599	0	0	0	1	0	0	0	0	18	0	5	3
10	454	1	6	0	0	0	0	2	6	16	0	20	6
11	1017	0	0	0	0	0	0	0	0	14	0	0	2
12	4	1	1	0	10	41	62	255	71	59	55	1	1
13	14	3	2	61	97	17	1	14	1	48	0	0	3
14	127	5	33	0	0	0	2	121	226	116	0	2	1
15	617	2	0	0	0	0	0	0	4	19	0	3	3
20 Clusters													
1	31	1	100	0	0	0	0	194	77	58	0	2	1
2	450	5	8	0	0	0	0	7	2	48	0	0	0
3	382	1	7	0	0	0	0	2	4	8	0	21	4

	Astro	Astro-M	Astro & GeoPhys	BioChem & MBio	BioChem & MBio-M	BioTech & AMBio-M	EnvSc & Eco-M	GeoChem & GeoPhys	GeoChem & GeoPhys-M	Multidisciplinary	Oceanography	Physics	Physics-M
4	861	0	0	0	0	0	0	0	0	9	0	0	0
5	283	15	0	0	0	0	0	13	21	28	0	29	109
6	560	4	0	0	0	0	0	3	1	14	0	3	8
7	449	8	9	0	0	0	0	3	1	47	0	4	0
8	713	0	0	0	0	0	0	0	0	15	0	2	1
9	574	0	0	0	1	0	0	0	0	16	0	2	3
10	323	3	2	0	0	0	0	26	49	100	0	0	1
11	378	3	0	0	0	0	0	0	4	11	0	1	4
12	7	2	1	61	103	53	22	24	0	59	2	0	2
13	31	1	159	0	0	0	0	282	17	53	0	1	0
14	200	12	11	0	3	0	1	35	34	70	0	0	1
15	313	2	13	0	0	0	0	6	32	25	0	1	0
16	222	1	40	0	0	0	0	79	3	93	0	15	3
17	119	5	3	0	0	0	2	65	174	85	0	2	0
18	352	2	10	0	0	0	0	2	2	47	0	2	0
19	665	0	0	0	0	0	0	0	0	7	0	0	1
20	1	1	1	0	2	5	41	237	70	37	53	1	0

Table 6: Clustering abstracts in the *conflated_SC_sampled* dataset with 5 clusters.

	Astro	Astro-M	Astro & GeoPhys	BioChem & MBio	BioChem & MBio-M	BioTech & AMBio-M	EnvSc & Eco-M	GeoChem & GeoPhys	GeoChem & GeoPhys-M	Multidisc-iplinary	Oceano-graphy	Physics	Physics-M
Trial 1													
1	0.05	0.00	0.00	0.00	0.00	0.00	0.31	0.02	0.01	0.02	0.00	0.76	0.99
2	0.00	0.00	0.00	0.00	0.09	0.93	0.68	0.14	0.09	0.05	1.00	0.00	0.00
3	0.89	0.98	0.10	0.00	0.01	0.00	0.00	0.07	0.25	0.56	0.00	0.17	0.01
4	0.00	0.00	0.00	1.00	0.90	0.07	0.00	0.00	0.00	0.04	0.00	0.00	0.00
5	0.06	0.01	0.90	0.00	0.00	0.00	0.01	0.77	0.65	0.33	0.00	0.07	0.00
Trial 2													
1	0.04	0.00	0.88	0.00	0.00	0.00	0.00	0.69	0.21	0.24	0.00	0.08	0.00
2	0.61	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.10	0.00	0.90	1.00
3	0.00	0.00	0.00	0.00	0.09	0.95	0.99	0.14	0.08	0.05	1.00	0.00	0.00
4	0.35	0.99	0.12	0.00	0.01	0.00	0.01	0.16	0.70	0.56	0.00	0.03	0.00
5	0.00	0.00	0.00	1.00	0.90	0.05	0.00	0.01	0.00	0.04	0.00	0.00	0.00
Trial 3													
1	0.33	0.99	0.07	0.00	0.01	0.00	0.34	0.09	0.31	0.47	0.00	0.00	0.03
2	0.05	0.00	0.92	0.00	0.00	0.00	0.00	0.79	0.63	0.34	0.00	0.04	0.00
3	0.00	0.00	0.00	0.00	0.09	0.96	0.66	0.11	0.06	0.05	1.00	0.00	0.00
4	0.00	0.00	0.00	1.00	0.90	0.04	0.00	0.00	0.00	0.04	0.00	0.00	0.00
5	0.62	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.11	0.00	0.95	0.97

3

Table 7: Clustering abstracts in the *conflated_SC_sampled* dataset with 10 clusters.

	Astro	Astro-M	Astro & GeoPhys	BioChem & MBio	BioChem & MBio-M	BioTech & AMBio-M	EnvSc & Eco-M	GeoChem & GeoPhys	GeoChem & GeoPhys-M	Multidisc- iplinary	Oceano- graphy	Physics	Physics-M
Trial 1													
1	0.03	0.00	0.82	0.00	0.00	0.00	0.00	0.55	0.07	0.20	0.00	0.00	0.00
2	0.00	0.00	0.00	1.00	0.87	0.01	0.00	0.00	0.00	0.03	0.00	0.00	0.00
3	0.08	0.78	0.01	0.00	0.02	0.00	0.00	0.04	0.05	0.13	0.00	0.02	0.07
4	0.06	0.17	0.12	0.00	0.00	0.00	0.01	0.19	0.70	0.24	0.00	0.00	0.00
5	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.14	0.90
6	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.02	0.00	0.00	0.19	0.00	0.00
7	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.81	0.02
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.12	0.04	0.81	0.00	0.00
9	0.78	0.05	0.05	0.00	0.00	0.00	0.00	0.02	0.06	0.32	0.00	0.03	0.01
10	0.00	0.00	0.00	0.00	0.11	0.99	0.11	0.00	0.00	0.02	0.00	0.00	0.00
Trial 2													
1	0.02	0.00	0.09	0.00	0.00	0.00	0.00	0.27	0.71	0.18	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.87	0.02	0.00	0.00	0.19	0.00	0.00
3	0.53	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.76	0.02
4	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.19	0.97
5	0.00	0.00	0.00	0.00	0.11	0.99	0.13	0.00	0.00	0.01	0.00	0.00	0.00
6	0.04	0.72	0.01	0.00	0.03	0.00	0.00	0.03	0.03	0.07	0.00	0.00	0.01
7	0.35	0.27	0.10	0.00	0.00	0.00	0.00	0.04	0.11	0.38	0.00	0.03	0.00
8	0.02	0.00	0.79	0.00	0.00	0.00	0.00	0.52	0.06	0.18	0.00	0.02	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.09	0.03	0.81	0.00	0.00
10	0.00	0.00	0.00	1.00	0.86	0.01	0.00	0.00	0.00	0.03	0.00	0.00	0.00
Trial 3													
1	0.00	0.00	0.00	0.00	0.00	0.00	0.59	0.02	0.00	0.00	0.00	0.00	0.00
2	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.16	0.96
3	0.07	0.96	0.03	0.00	0.02	0.00	0.00	0.03	0.05	0.13	0.00	0.01	0.01
4	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.03	0.02	0.01	1.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.13	0.99	0.12	0.01	0.00	0.02	0.00	0.00	0.00
6	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.80	0.01
7	0.82	0.04	0.07	0.00	0.00	0.00	0.00	0.02	0.08	0.36	0.00	0.02	0.01
8	0.03	0.00	0.08	0.00	0.00	0.00	0.00	0.34	0.78	0.23	0.00	0.00	0.00
9	0.03	0.00	0.83	0.00	0.00	0.00	0.00	0.53	0.06	0.20	0.00	0.02	0.00
10	0.00	0.00	0.00	1.00	0.84	0.01	0.00	0.00	0.00	0.03	0.00	0.00	0.00

Table 8: Clustering abstracts in the *conflated_SC_sampled* dataset with 15 clusters.

ography

	Astro	Astro-M	Astro & GeoPhys	BioChem & MBio	BioChem & MBio-M	BioTech & AMBio-M	EnvSc & Eco-M	GeoChem & GeoPhys	GeoChem & GeoPhys-M	Multidisc-iplinary	Oceano-graphy	Physics	Physics-M
Trial 1													
1	0.00	0.00	0.00	0.58	0.31	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
2	0.03	0.00	0.64	0.00	0.00	0.00	0.00	0.46	0.05	0.16	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.12	0.99	0.06	0.00	0.00	0.01	0.00	0.00	0.00
4	0.02	0.66	0.01	0.00	0.02	0.00	0.00	0.01	0.01	0.04	0.00	0.00	0.01
5	0.06	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.04	0.00	0.46	0.10
6	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.85
7	0.01	0.00	0.22	0.00	0.00	0.00	0.00	0.28	0.54	0.15	0.00	0.00	0.00
8	0.00	0.00	0.00	0.42	0.55	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.44	0.02	0.00	0.00	0.00	0.00	0.00
10	0.10	0.20	0.03	0.00	0.00	0.00	0.00	0.07	0.29	0.24	0.00	0.00	0.00
11	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.45	0.03
12	0.54	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.03	0.01
13	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.02	0.00	0.00	0.20	0.00	0.00
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.10	0.03	0.80	0.00	0.00
15	0.21	0.14	0.08	0.00	0.00	0.00	0.00	0.01	0.00	0.21	0.00	0.03	0.00
Trial 2													
1	0.00	0.00	0.00	0.24	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.44	0.03	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.51	0.27	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
4	0.03	0.01	0.08	0.00	0.00	0.00	0.00	0.25	0.73	0.22	0.00	0.00	0.00
5	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.17	0.96
6	0.00	0.00	0.00	0.09	0.36	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.07	0.34	0.24	0.02	0.00	0.03	0.01	0.00	0.00
8	0.04	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.79	0.01
9	0.00	0.00	0.00	0.16	0.04	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.06	0.65	0.01	0.00	0.00	0.00	0.00	0.03	0.04	0.10	0.00	0.00	0.02
11	0.16	0.33	0.10	0.00	0.00	0.00	0.00	0.01	0.00	0.17	0.00	0.03	0.00
12	0.02	0.00	0.80	0.00	0.00	0.00	0.00	0.53	0.06	0.18	0.00	0.01	0.00
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.09	0.02	0.80	0.00	0.00
14	0.00	0.00	0.00	0.00	0.01	0.21	0.31	0.00	0.00	0.00	0.20	0.00	0.00
15	0.66	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.07	0.22	0.00	0.01
Trial 3													
1	0.00	0.00	0.00	0.00	0.12	0.99	0.06	0.00	0.00	0.01	0.00	0.00	0.00
2	0.08	0.00	0.04	0.00	0.00	0.00	0.00	0.02	0.02	0.06	0.00	0.32	0.13
3	0.61	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.01
4	0.10	0.06	0.02	0.00	0.00	0.00	0.01	0.07	0.31	0.26	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.21	0.00	0.00
6	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.85

	Astro	Astro-M	Astro & GeoPhys	BioChem & MBio	BioChem & MBio-M	BioTech & AMBio-M	EnvSc & Eco-M	GeoChem & GeoPhys	GeoChem & GeoPhys-M	Multidisc-iplinary	Oceano-graphy	Physics	Physics-M
7	0.02	0.00	0.61	0.00	0.00	0.00	0.00	0.46	0.03	0.16	0.00	0.00	0.00
8	0.02	0.00	0.23	0.00	0.00	0.00	0.00	0.27	0.54	0.15	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.42	0.02	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.46	0.71	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.09	0.03	0.79	0.00	0.00
12	0.03	0.56	0.01	0.00	0.02	0.00	0.00	0.01	0.01	0.04	0.00	0.00	0.00
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.62	0.01
14	0.00	0.00	0.00	0.54	0.14	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
15	0.12	0.37	0.09	0.00	0.00	0.00	0.00	0.01	0.00	0.15	0.00	0.02	0.00

Table 9: Clustering abstracts in the *conflated_SC_sampled* dataset with 20 clusters.

	Astro	Astro-M	Astro & GeoPhys	BioChem & MBio	BioChem & MBio-M	BioTech & AMBio-M	EnvSc & Eco-M	GeoChem & GeoPhys	GeoChem & GeoPhys-M	Multidisc-iplinary	Oceano-graphy	Physics	Physics-M
Trial 1													
1	0.01	0.00	0.23	0.00	0.00	0.00	0.00	0.25	0.08	0.07	0.00	0.00	0.00
2	0.10	0.05	0.05	0.00	0.00	0.00	0.00	0.07	0.08	0.18	0.00	0.00	0.08
3	0.00	0.00	0.00	0.00	0.00	0.00	0.42	0.01	0.00	0.00	0.00	0.00	0.00
4	0.02	0.00	0.10	0.00	0.00	0.00	0.00	0.05	0.29	0.06	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.35	0.00	0.00	0.00	0.21	0.00	0.00
6	0.00	0.00	0.00	0.07	0.21	0.03	0.00	0.00	0.00	0.02	0.00	0.00	0.00
7	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.01
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.08	0.03	0.53	0.00	0.00
9	0.14	0.14	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.04	0.00
10	0.00	0.00	0.00	0.34	0.25	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	0.02	0.03	0.01	0.00	0.00	0.00	0.01	0.09	0.37	0.09	0.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.03	0.96	0.06	0.00	0.00	0.01	0.00	0.00	0.00
13	0.00	0.00	0.00	0.59	0.51	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
14	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.35	0.06
15	0.01	0.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
16	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.59	0.01
17	0.02	0.00	0.56	0.00	0.00	0.00	0.00	0.39	0.02	0.12	0.00	0.00	0.00
18	0.12	0.21	0.01	0.00	0.00	0.00	0.00	0.02	0.08	0.18	0.00	0.00	0.00
19	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.84
20	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.03	0.00	0.01	0.26	0.00	0.00
Trial 2													
1	0.00	0.00	0.00	0.19	0.18	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.02	0.00	0.76	0.00	0.00	0.00	0.00	0.51	0.05	0.17	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.02	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.16	0.21	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.61
6	0.10	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.05	0.00	0.15	0.16
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.23
8	0.01	0.62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
9	0.00	0.00	0.00	0.09	0.36	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
10	0.00	0.00	0.00	0.35	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.00	0.01	0.32	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.17	0.00	0.00
13	0.18	0.13	0.02	0.00	0.00	0.00	0.00	0.06	0.21	0.30	0.00	0.00	0.00
14	0.00	0.00	0.00	0.21	0.10	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00
15	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.68	0.00
16	0.63	0.04	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.02	0.01
17	0.00	0.00	0.00	0.00	0.08	0.96	0.08	0.00	0.00	0.01	0.00	0.00	0.00
18	0.01	0.00	0.03	0.00	0.00	0.00	0.01	0.27	0.53	0.16	0.00	0.00	0.00

	Astro	Astro-M	Astro & GeoPhys	BioChem & MBio	BioChem & MBio-M	BioTech & AMBio-M	EnvSc & Eco-M	GeoChem & GeoPhys	GeoChem & GeoPhys-M	Multidisc-iplinary	Oceano-graphy	Physics	Physics-M
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.01	0.50	0.00	0.00
20	0.03	0.20	0.14	0.00	0.00	0.00	0.00	0.05	0.17	0.07	0.00	0.00	0.00
Trial 3													
1	0.00	0.00	0.00	0.42	0.30	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.70
3	0.00	0.00	0.00	0.00	0.00	0.00	0.41	0.02	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.35	0.26	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.04	0.19	0.01	0.00	0.05	0.00	0.01	0.04	0.01	0.12	0.00	0.00	0.00
6	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.43	0.01
7	0.14	0.07	0.11	0.00	0.00	0.00	0.00	0.01	0.00	0.16	0.00	0.04	0.00
8	0.01	0.00	0.25	0.00	0.00	0.00	0.00	0.16	0.34	0.08	0.00	0.00	0.00
9	0.00	0.00	0.00	0.24	0.32	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
10	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.06	0.28
11	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.20	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.80	0.00	0.00
13	0.17	0.01	0.02	0.00	0.00	0.00	0.00	0.07	0.29	0.29	0.00	0.00	0.00
14	0.00	0.32	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
15	0.02	0.00	0.59	0.00	0.00	0.00	0.00	0.44	0.03	0.14	0.00	0.01	0.00
16	0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.02	0.01
17	0.00	0.00	0.00	0.00	0.08	0.98	0.06	0.00	0.00	0.01	0.00	0.00	0.00
18	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.22	0.30	0.07	0.00	0.00	0.00
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.43	0.00
20	0.01	0.42	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00

Copyright Transfer Statement

Publisher: Akadémiai Kiadó Zrt., Budapest, Hungary

Author

Name: Michael Gowanlock

Address: 1680 East-West Road, University of Hawaii, Post Building, Honolulu, HI, 96822

E-mail address: gowanloc@hawaii.edu

Article information

Title: Assessing Researcher Interdisciplinarity: A Case study of the University of Hawaii NASA Astrobiology Institute

Journal title: **Scientometrics**

Co-authors: Rich Gazan

I. Transfer of copyright

By execution of the present Statement Author transfers copyright and assigns exclusively to Publisher all rights, title and interest that Author may have (for the extent transferable) in and to the Article and any revisions or versions thereof, including but not limited to the sole right to print, publish and sell the Article worldwide in all languages and media. Transfer of the above rights is referred to as those of the final and published version of the Article but does not restrict Author to self-archive the preprint version of his/her paper (see Section III).

II. Rights and obligations of Publisher

The Publisher's rights to the Article shall especially include, but shall not be limited to:

- ability to publish an electronic version of the Article via the website of the publisher Akademiai Kiado, www.akademiai.com (in Hungary), as well as the co-publisher's website, www.SpringerLink.com (outside of Hungary) or any other electronic format or means of electronic distribution provided by or through Akademiai Kiado or Springer from time to time, selling the Article world-wide (through subscriptions, Pay-per-View, single archive sale, etc.)
- transforming to and selling the Article through any electronic format
- publishing the Article in the printed Journals as listed on the official Website of Publisher
- transferring the copyright and the right of use of the Article onto any third party
- translating the Article
- **taking measures on** behalf of the Author against infringement, inappropriate use of the Article, libel or plagiarism

Publisher agrees to send the text of the Article to the e-mail address of Author indicated in the present Statement for preview before the first publishing either in paper and/or electronic format (Proof). Author shall return the corrected text of the Article within 2 days to the Publisher. Author shall, however, not make any change to the content of the Article during the First Proof preview.

III. Rights and obligations of Author

The Author declares and warrants that he/she is the exclusive author of the Article – or has the right to represent all co-authors of the Article (see Section IV) – and has not granted any exclusive or non-exclusive right to the Article to any third party prior to the execution of the present Statement and has the right therefore to enter into the present Statement and entitle the Publisher the use of the Article subject to the present Statement. By executing the present Statement Author confirms that the Article is free of plagiarism, and that Author has exercised reasonable care to ensure that it is accurate and to the best of Author's knowledge does not contain anything which is libelous or obscene or infringes on anyone's copyright right of privacy or other rights. The Author expressly acknowledges and accepts that he she shall be entitled to no royalty or any other fee related to any use of the Article subject to the present Statement. The Author further accepts that he/she shall not be entitled to dispose of the copyright of the final published version of the Article or make use of this version of the Article in any manner after the execution of the present Statement. The Author is entitled, however, to self-archive the preprint version of his/her manuscript. The preprint version is the Author's manuscript or the galley proof or the Author's manuscript along with the corrections made in the course of the peer review process. The Author's right to self-archive is irrespective of the format of the preprint (.doc, .tex., .pdf) version and self-archiving includes the free circulation of this file via e-mail or publication of this preprint on the Author's webpage or on the Author's institutional repository with open or restricted access. When self-archiving a paper the Author should clearly declare that the archived file is not the final published version of the paper, he/she should quote the correct citation and enclose a link to the published paper (http://dx.doi.org/[DOI of the Article without brackets]).

IV. Use of third party content as part of the Article

When not indicating any co-authors in the present Statement Author confirms that he/she is the exclusive author of the Article. When indicating co-authors in the present Statement Author declares and warrants that all co-authors have been listed and Author has the exclusive and unlimited right to represent all the co-authors of the Article and to enter into the present Statement on their behalf and as a consequence all declarations made by Author in the present Statement are made in the name of the co-authors as well. Author also confirms that he/she shall hold Publisher harmless of all third-party claims in connection to non-authorized use of the Article by Publisher. Should Author wish to reuse material sourced from third parties such as other copyright holders publishers authors etc. as part of the Article Author bears responsibility for acquiring and clearing of the third party permissions for such use before submitting the Article to the Publisher for acceptance. Author shall hold Publisher harmless from all third party claims in connection to the unauthorized use of any material under legal protection forming a part of the Article

V. Other provisions

Subject to the present Statement the Article shall be deemed as first published within the Area of the Hungarian Republic. Therefore the provisions of the Hungarian law especially the provisions of Act LXXVI of 1999 on Copyrights shall apply to the rights of the Parties with respect to the Article. For any disputes arising from or in connection with the present Statement Parties agree in the exclusive competence of the Central District Court of Pest or the Capital Court of Budapest respectively

Date and place of signature

Jan. 12, 2012