The Spitzer Bibliography Database: Bibliographic Statistics

Elena Scire[*a], Ben Hiu Pan Chan[b], Nancy Silbermann[a], Aomawa Shields[c]
[a] Spitzer Science Center, California Institute of Technology, MS 220-6, Pasadena, CA 91125; [b] IPAC/NED, California Institute of Technology, Mail Code 100-22, Pasadena, CA, 91125; [c] Department of Astronomy, University of Washington, Box 351580, Seattle, WA 98195-1580

1. ABSTRACT

The Spitzer Science Center maintains a database of peer-refereed publications utilizing observations made by the Spitzer Space Telescope[5]. Originally intended as a way to easily track these publications with limited resources, the database has grown in scope to provide more services for investigators. The design and population of the system and some interesting insights into the use of Spitzer data are presented.

2. INTRODUCTION

The Spitzer Bibliographical Database (SBD) (http://sohelp2.ipac.caltech.edu/bibsearch/) was originally conceived as a tool measure publication rates for the 2008 NASA Senior Review. Before the database existed, information on Spitzer publications was kept on static web pages, which were manually updated on a monthly basis. In preparation for the Senior Review the Infrared Processing and Analysis Center (IPAC) produced a list of hits of probable Spitzer papers using a journal search engine developed at IPAC (see section 3.3) that searched the major astronomical journals: Astronomy & Astrophysics (A&A), Astronomical Journal (AJ), Astrophysics and the Astrophysics Supplemental Series (ApJ and ApJS) and Monthly Notices of the Royal Astronomical Society (MNRAS). These journals were picked because together they make up 94% of the total publications that use Spitzer data. This list of articles needed to be cross-checked against the list of known Spitzer papers from the website. The remaining papers then needed to be checked to see if the authors had used Spitzer data in the paper. In order to facilitate this the Spitzer Bibliographical Database was born.

After the 2008 Senior Review an automated search script was written that runs against the SAO/NASA Astrophysics Data System (ADS)[6] weekly, and the database continued to grow. Since the information was stored in a SQL database, it was easy to query and other uses for the data were found. In particular, in the summer of 2008 an effort was made to go through the articles that already existed in the database and match them up with the observations used in the articles. Once this had been done, a web interface for the database was written and the database was made available for use by the general astronomical community in December 2008.

3. DESIGN

The core of the SBD is a MySql database surrounded by perl and shell scripts. The web interface is written in PHP. These technologies were chosen because they are robust, platform independent and open source.

The SBD allows searches by Program Identifier number (PID) and Astronomical Observation Request (AOR) number, Legacy Program, Title, Author, Date, Instrument, and Paper Type (Figure 1). In addition, on the results page the search criteria are listed so that a user who saves a search can easily tell which parameters were used (Figure 2). For the title and author searches, the search terms are highlighted with red font in the search results.

The design of the search form and the results page were influenced by the Hubble[7], Chandra[8] and ADS search interfaces.

[*] escire@ipac.caltech.edu, phone: +1 (626) 395-3315

Observatory Operations: Strategies, Processes, and Systems III, edited by
David R. Silva, Alison B. Peck, B. Thomas Soifer, Proc. of SPIE Vol. 7737, 77371V
© 2010 SPIE · CCC code: 0277-786X/10/$18 · doi: 10.1117/12.857735

3.1 Database

The MySql database has five tables: pending, rejected, valid, AOR and PID. The pending table is where all unprocessed articles are stored. The rejected table stores papers that have been determined to not use Spitzer data directly. The valid table contains all papers that use Spitzer data directly. The AOR and PID tables contain the mappings between the papers and the observations the authors used. We are using a mirrored database system. The staging database is used to compile the results, and all the scripts described below are run against it. The live database is a read-only copy that users interact with through the PHP query form.

The database is populated in three steps. The first step is a weekly search on titles and abstracts using ADS. The second step is a monthly full-text search using the journals' websites. The third step is a yearly run against the IPAC home grown full-text search engine. The three steps are outlined below:

3.2 ADS searches (Step 1)

ADS searches are almost completely automated. A Unix cron job executes the controlling shell script which does the following:

1. Connects to ADS and runs a search on certain search terms using the Astro::ADS perl library [9]
2. Downloads the results of the search
3. Parses the result set into individual articles
4. Checks to see if each article already exists in the database. If not, the article is inserted into the pending table. This check is run by comparing the article's ADS bibliographic identifier code (bibcode) [10]. If the ADS bibcode is missing, a second check is run on the uniqueness of the journal, volume and page values.
5. Creates a list of new articles

Once the list of new articles is generated, they are then manually checked and linked to Spitzer observations they use. A second shell script is used to perform the following:

1. Ingests the results of the manual check into the database
2. Creates a backup of the database
3. Creates a copy of the staging database for transfer to the live database.

We have found that the ADS searches produce ~80% of the articles in the database. In general this is because authors are very good about specifying how their data was acquired in the abstract or the acknowledgement statement, both of which appear in ADS.

3.3 Full Text Searches (Steps 2 & 3)

Many journals have full-text search engines built into their websites. A second library of scripts was developed to parse the results of these search engines. These journals are published by: IOP Publishing (ApJ, ApJS, AJ, ect.), EDP Sciences (A&A) and Chicago Journals (Publications of the Astronomical Society of the Pacific (PASP)). Searches are run on these journals' websites manually on a monthly basis and the results downloaded and passed into the full text scripts which:

1. Parses the result set into individual articles
2. Checks to see if each article already exists in the database. If not, the article is inserted into the pending table.
3. Creates a list of new articles for manual checks

After the articles are checked the results are manually ingested back into the staging database. The new articles are propagated out to the live database during the next weekly ADS run. These searches usually find ~15% of the articles. In general the major thrust of these articles are not Spitzer related, however the authors use Spitzer data in a minor capacity, usually to provide additional evidence for their main argument.

In addition to these full text searches, a separate search is run on ApJ, ApJ, AJ, A&A and MNRAS articles by a journal search engine developed at IPAC. It accesses the on-line e-journals to locate and index phrases of interest from the full text content. This search engine was used to create the first version of the SBD. This second full text search often returns results where one of the keywords was found in an image or table caption, or a footnote. This search does recover all the articles found by the other two methods, but because it requires more resources to setup and execute it is run yearly. This final search finds ~20% of the articles from MNRAS and as much as 5% of the total articles for the other journals.

3.4 Database standardization

We are using information about journal articles from many sources, and they do not always have the same formatting. Articles found in the monthly full-text search never have an ADS bibcode, some articles have the title in all capital letters and the formatting of the authors can vary wildly. In order to standardize the formatting, a second set of scripts to query ADS was created. These scripts:

1. Take an input list of journal, volume and page values.
2. Query ADS for articles
3. Return a SQL statement to correct the article entry in the SBD by using the values in ADS.

These SQL statements are checked by hand for accuracy before being executed.

4. JOURNAL ARTICLE CRITERIA

We include only refereed journal articles that have a permanent ADS URL and use Spitzer data as a primary resource or are about the telescope, instruments or software developed at the Spitzer Science Center. In some cases, we also keep track of theoretical articles if they pertain primarily to Spitzer data. We do not track pre-prints or errata.

We search on the name of the telescope, instruments, Legacy Science Programs and Exploration Science Programs. Some Legacy program names were omitted from the search because they produce far more false positives then actual articles (SIMPLE, GOALS, Taurus, Cygnus-X).

We originally did not track papers that use catalogues of sources and mosaics produced by the Legacy Science Programs since this data is used as a secondary source, and the paper can be found by looking at the articles that reference the Legacy Science Programs' seminal article. Part of the Legacy Science catalogues are also served by IRSA [11], and therefore are not under the control of the Spitzer Science Center. Tracking for these papers was added to the SBD in Spring 2010 because a higher fraction of observers now use the enhanced Legacy Science Program data products, rather then the original Legacy data.

When an article is found that satisfies the criteria, an effort is made to match it to the data used in the paper by the authors. This is done by hand using information given in the article to produce a list of Spitzer observations that may have been used by the authors. In some cases the authors provide excellent information and matching to observations is easy. In cases where it was impossible to pinpoint the exact observations that were used to perform the data analysis we erred on the side of caution and linked all the possible observations to the paper. In the case of the Legacy and Exploration Science programs where most or all of the data was used, the paper was only linked to the PID and not the AOR number, as some of these programs contain thousands of AORs. This is especially true in the case of the deep fields, as they contain many AORs that layer over each other and so a search on specific coordinates returns the entire program anyway. (GOODS, COSMOS, ELAIS N and S, ect…).

As part of the Senior Review process, we also produced a count of publications for the Hubble Space Telescope and the Chandra X-ray Observatory using the same method outlined above. We then ran a cross-check with the articles claimed by those observatories (using the databases given in section 3). Our count of Hubble's publications agrees to within 5% of Hubble's count [1] but our count of Chandra's publications is on average 112 papers/year higher then Chandra's count. We believe this is because Chandra classifies their papers using a more granular system and it is harder to produce a list of articles from their search engine that meet the same criteria we are using.

5. SPITZER BIBLOGRAPHICAL DATABASE USAGE

On average, the database has been used to run 68 searches a month since it was launched. As shown in Table 1, the most popular search type is by PID. This is expected behavior, because what makes the SBD useful to the community is the ability to check if a certain observation has already been published.

We also use the database internally to check for a publication when closing out grant contacts.

6. SPITZER PUBLICATION STATISTICS

Data from Spitzer's two imaging instruments (IRAC[2] and MIPS[4]) are more frequently published than data from the spectrograph (IRS[3]) (Figure 3). This is due in most part to the time elapsed between observation and publication. Figure 4 shows the time elapsed for Cycle 2 data between the release of the data to the PIs and the appearance of the first published paper containing the data. Cycle 2 observations began in June 2005. In general it takes 9-12 months for the first data to be published, and then it picks up quickly for IRAC, and more slowly for MIPS and IRS. This is most likely due to the complexity of analyzing any spectra, and the subtleties involved in reducing the two longest wavelengths of MIPS data. For this cycle 21% of the MIPS AORs, 31% of the IRAC AORs and 1.6% of the IRS AORs had a proprietary period less then the standard proprietary period of one year, so the lag in IRS and MIPS publications is not due to the proprietary period because this does not match the trend in Figure 4.

Figure 5 shows the percentage of publications by year using non-archival data (at least one author on the article matches a primary investigator (PI), technical contact (TC) or co-investigator (CoI) on the Spitzer program), archival data (no author matches the PI, TC or CoI) or a mixture of archival and non-archival data. The non-zero value of archival publications in 2004 is a direct result of the devotion of most of the observatory time during the first year to Legacy Science programs, which had no proprietary period and covered popular survey fields. The original Legacy Science programs (GOODS, GLIMPSE, SINGS, C2D, FEPS, SWIRE) remain among the most published Spitzer data (Table 2).

7. ACKNOWLDGEMENTS

This work is based [in part] on observations made with the Spitzer Space Telescope, which is operated by the Jet Propulsion Laboratory, California Institute of Technology under a contract with NASA.

8. REFERENCES

[1] Apai, D., Lagerstrom, J., Reid, I. N., Levay, K. L., Fraser, E., Nota, A., Henneken, E., "Lessons from a High-Impact Observatory: The Hubble Space Telescope's Science Productivity between 1998 and 2009", In preparation, (2010).
[2] Fazio, G. G., et al., "The Infrared Array Camera (IRAC) for the Spitzer Space Telescope", ApJS 154, 10 (2004).
[3] Houck, J., et al., "The Infrared Spectrograph on the Spitzer Space Telescope", ApJS 154, 18 (2004).
[4] Rieke, G., et al., "The Multiband Imaging Photometer for Spitzer", ApJS 154, 25 (2004).
[5] Werner, M., et al., "The Spitzer Space Telescope Mission", ApJS 154, 1 (2004).
[6] http://adsabs.harvard.edu/abstract_service.html
[7] http://archive.stsci.edu/hst/bibliography/
[8] http://cxc.harvard.edu/cgi-gen/cda/bibliography
[9] http://doc.adsabs.harvard.edu/abs_doc/help_pages/linking.html#Perl_Script_Access
[10] http://doc.adsabs.harvard.edu/abs_doc/help_pages/data.html#bibcodes
[11] http://irsa.ipac.caltech.edu/

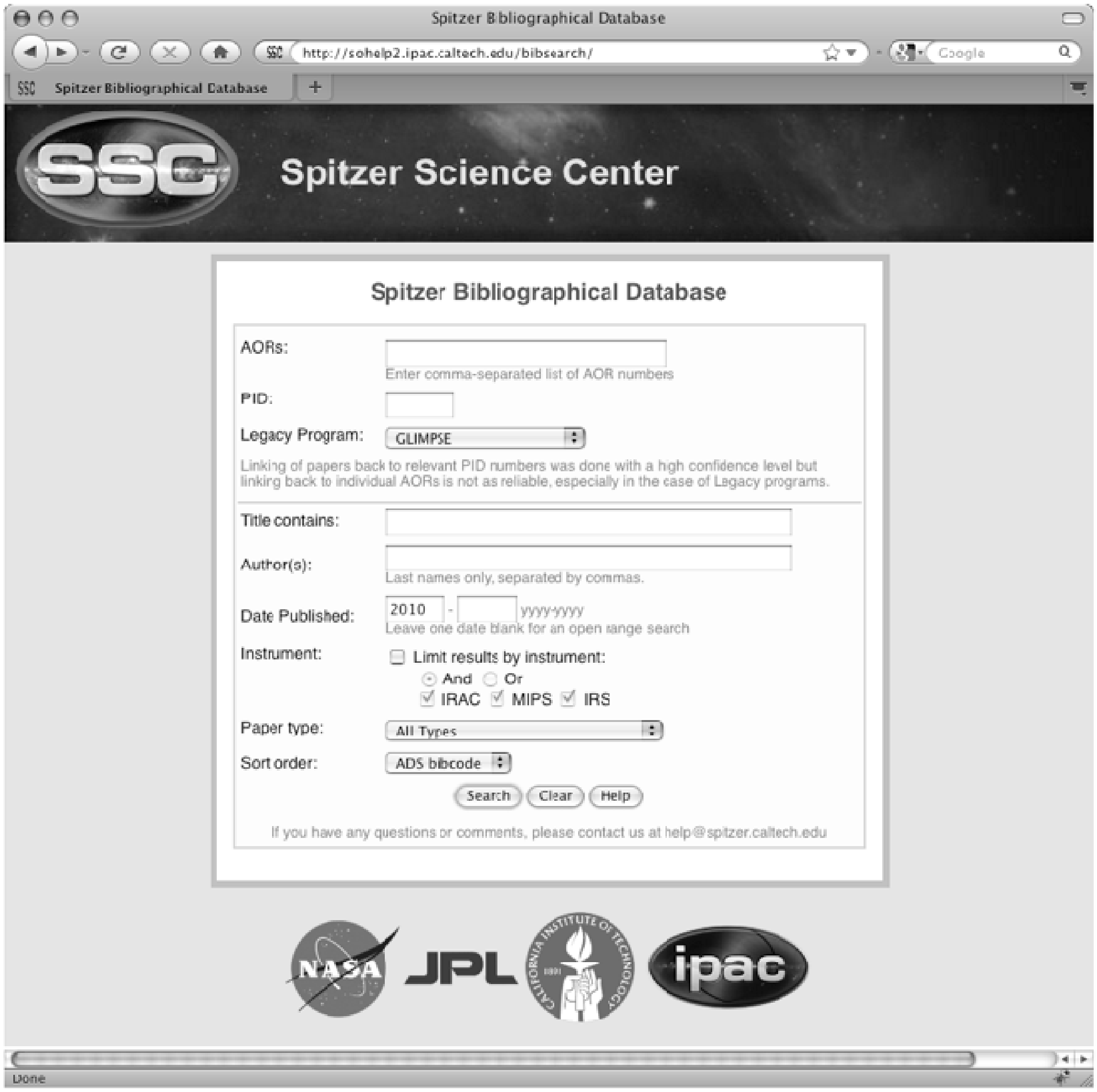

Figure 1: The Graphical User Interface (GUI) for the SBD

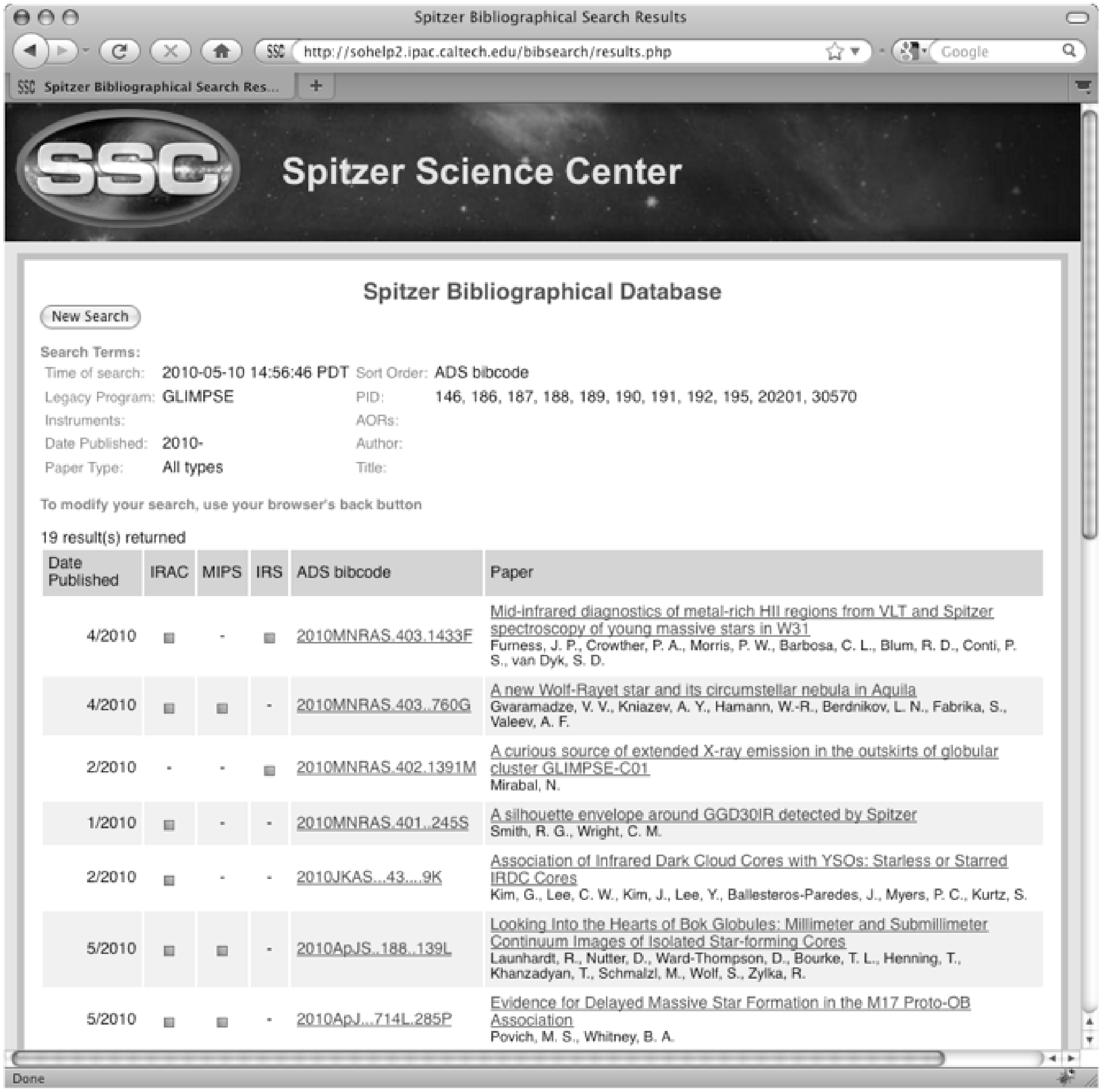

Figure 2: The results page for the SBD

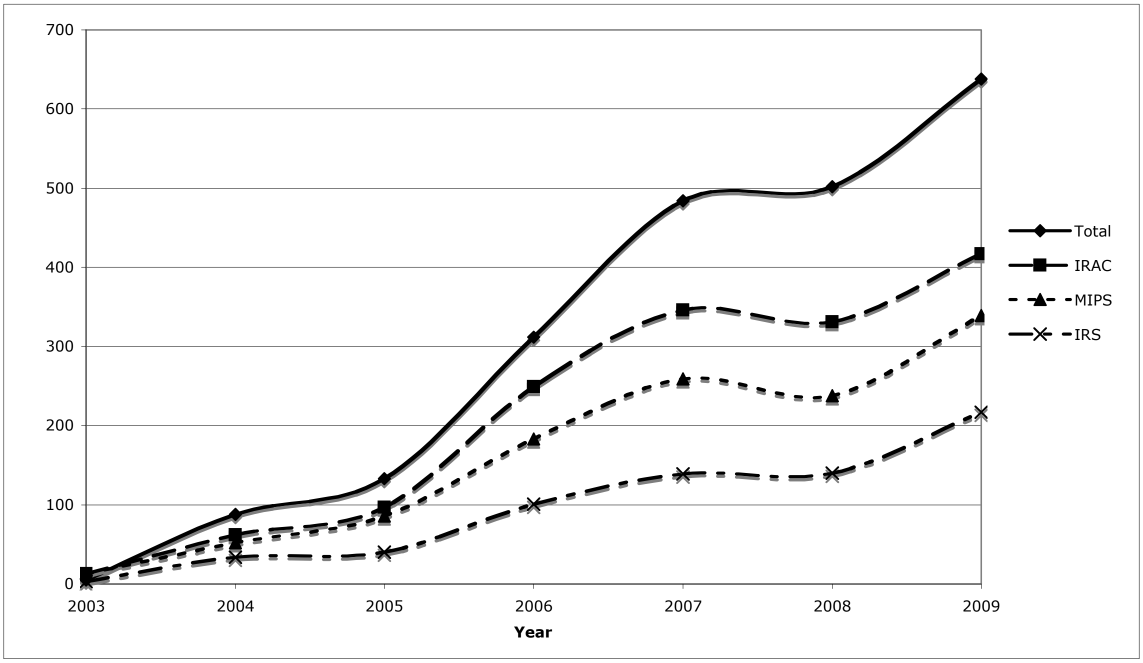

Figure 3: Spitzer articles by year, and articles containing data from the three Spitzer instruments (IRAC, IRS, MIPS) by year

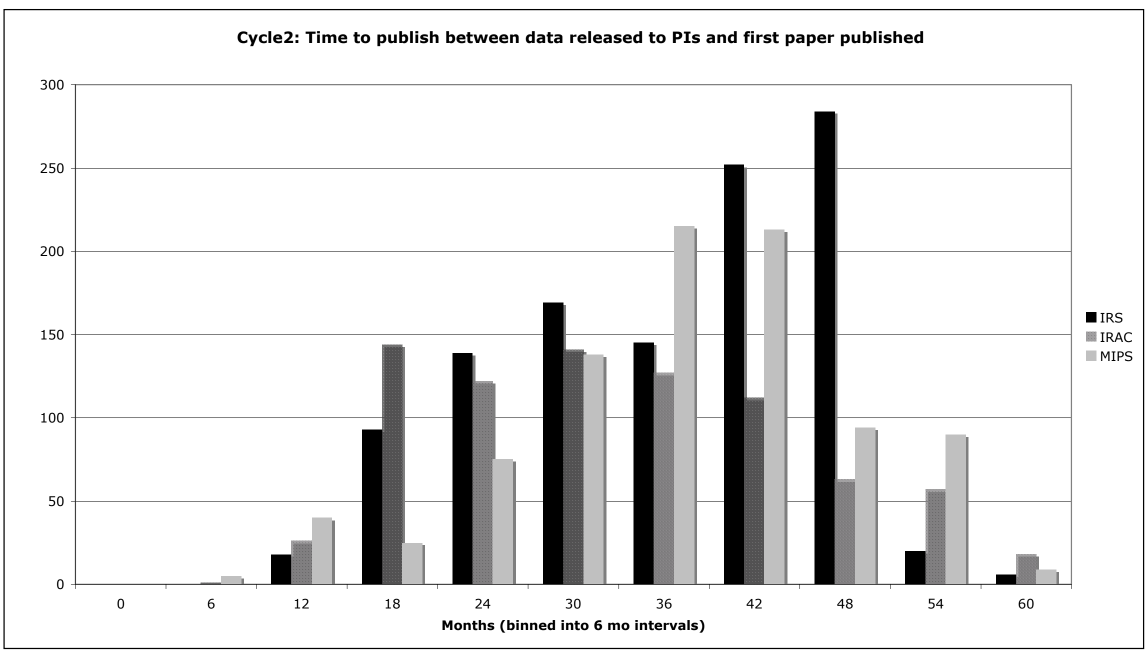

Figure 4: Months between the time the data was released to the PI and the publication of the first paper using the data. The date published used is the 'Month' field in ADS. The data has been binned into 6 month intervals.

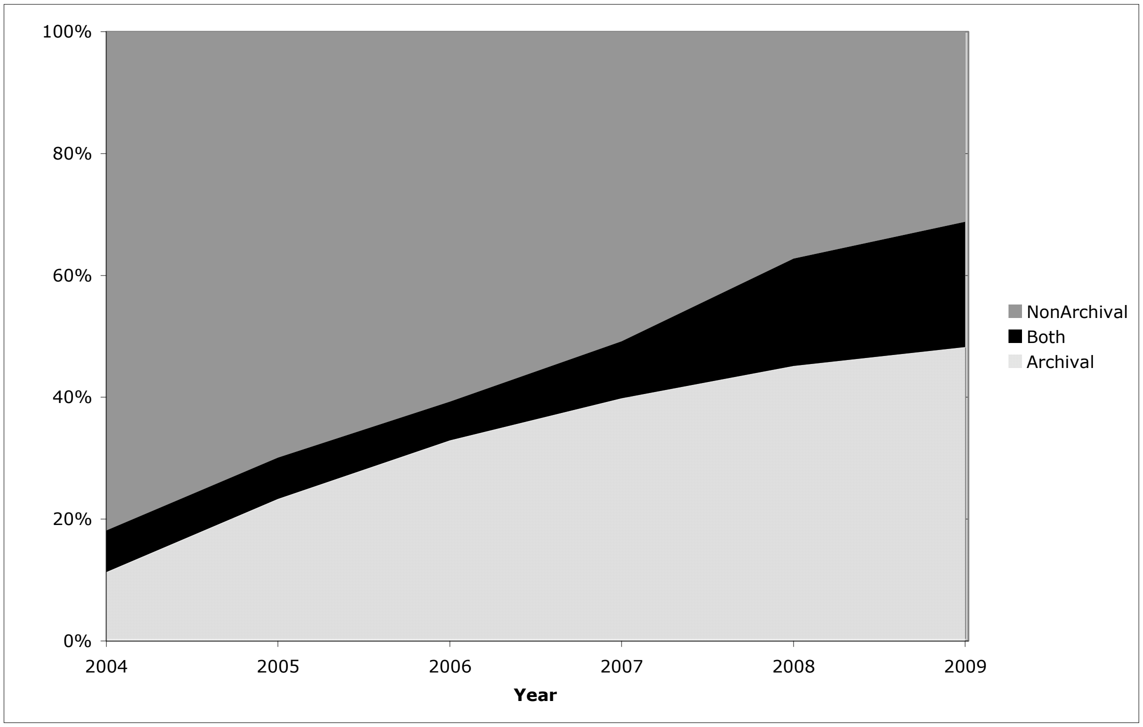

Figure 5: Percentage of publications that are archival, non-archival, or both by year.

Table 1: Searches run by category as a percentage of the total number of searches run.

Search Type	% of Total*
AOR	7.8
PID	41.4
Author	7.8
Title	17.2
Type	12.8
Year	21.6
Instrument	14.9

* Complex searches involving more then one search type can be run, ergo this column does not add up 100%

Table 2: Top twenty most published Spitzer programs to date

Program	Program Type	PIDs	# Publications
GLIMPSE	Legacy	146,186,187,188, 189,190,191,192, 195,20201,30570	208
C2D	Legacy	139,172,173,174, 175,176,177,178,179	169
GOODS	Legacy	169,194	153
SINGS	Legacy	159,193	142
SWIRE	Legacy	142,181,182,183,184	107

IRS_DISKS	GTO	2	86
MIPSGAL	Legacy	20597,30594	67
FLS_EXTRAGAL	DDT	26	61
DISKS_EVOLUTION	GTO	37	58
M_RIEKE_SURVEYS	GTO	81	57
PMYERS_CLUSTERS	GTO	6	56
YOUNG_LADA_DISKS	GTO	58	54
DWW_AGN	GTO	14	50
IRAC-SHALLOW	GTO	30	44
MIDIRATLAS	GTO	69	42
FEPS	Legacy	148	40
LA_ULIRGS2	GTO	105	39
SAGE	Legacy	20203	37
GS_SURVEY	GTO	8	35
MWWGTO-XRAYAGN	GTO	86	34