

**Title:** “Cosmic Archaeology: The Recent History of the Harvard-Smithsonian Center for Astrophysics Analyzed through Literature Heat Maps”

**Author:** McKenna E. Kardish

**Contributors:** Goodman, Alyssa; Pepe, Alberto; Beaumont, Chris; and Robert Simpson

**Affiliations:** Harvard-Smithsonian Center for Astrophysics

### **Abstract**

The purpose of this project is split into two parts- the first of which is testing the functionality of the Astrophysics Data System All-Sky Survey (ADSASS), using Aladin and WorldWide Telescope (WWT) to display heat maps showing where and why the Sky has been studied and by whom, over time. These tests are significant, as this project is, as far as we know, the first of its kind to use the ADSASS, or any heat map, to study the distribution of astronomy research projects on the Sky. The second part of the project focuses on utilizing the ADSASS to analyze the history of the output of astronomical articles specifically from the Harvard-Smithsonian Center for Astrophysics (CfA) through heat maps of article density portrayed on the sky through the articles’ objects’ coordinates. This study demonstrates if and how CfA scientists’ interests varied from the general interests of the Astrophysics community at various points from 2000-2012. Conducting the research also lead to identification of areas of improvement and changes for the ADSASS, Aladin, and WWT teams. The study concludes that once planned improvements to the “Select Tool” are implemented, analysis of the ADSASS heat maps will offer a very useful technique for studying and analyzing the history of astrophysics research at any institution, especially for the years 2000-2012 in which the collection of articles is the most complete.

## 1. Introduction

The Harvard-Smithsonian Center for Astrophysics (CfA) includes one of the oldest observatories in the nation, the Harvard College Observatory (HCO). The HCO was founded in 1839 and quickly became well-known in the realm of astrophysics, producing many significant discoveries and contributions including the Harvard Revised Photometry Catalogue, which gave rise to the HR star catalogue, and Henrietta Leavitt's discovery of the use of variable stars in calculating distances to astronomical objects. In 1955, the Smithsonian Astrophysical Observatory (SAO) moved to Cambridge and its relationship to the HCO was formalized in 1973, officially forming a joint venture named the Harvard-Smithsonian Center for Astrophysics (<http://www.cfa.harvard.edu>).

Today, the CfA is arguably one of the largest and most diverse astrophysical institutions in the world with the combined staff now over 300 scientists, studying a broad range of topics in astronomy, astrophysics, and earth and space sciences. Nearly 30,000 articles have been published by scientists working at the HCO and CfA since 1839 (<http://www.cfa.harvard.edu>).

With this in mind, we decided to study and analyze the immense history of the CfA through the articles published at the CfA by using the Astrophysics Data System All-Sky Survey (ADSASS). Because the CfA is such a dominant institution for the field of astrophysics, studying its past can tell us about its strengths as an institution as well as highlight its achievements.

From this broad goal, we began to narrow our focus to develop a 2-part project. The first part of the project aims to study the history of the CfA using the ADSASS. Using some of the 30,000 scientific papers and articles published from the CfA, we will analyze the CfA's history by asking questions such as what objects are more interesting to Harvard than to the rest of the world or vice versa, when were these objects studied, and does Harvard display trends in its

research interests? As of the moment, no study using analytic, geo-spatial tools has ever been conducted delving into Harvard's past to answer these questions. We decided to use the ADSASS to analyze the distribution of articles among objects as well as the time scale in order to attempt to answer these questions.

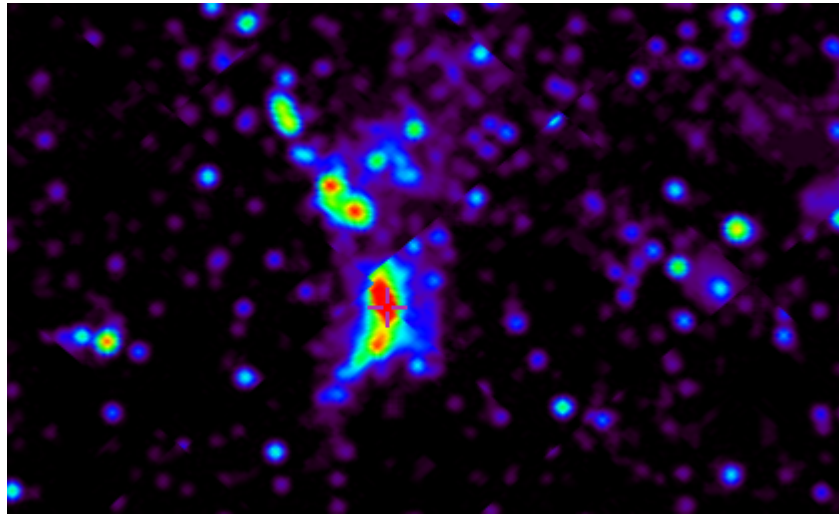
The second part of the project focused on testing the ADSASS and developing strategies for improvement of the program as the project heavily relied on the ADSASS for analyzing our data. The All-Sky Survey is a relatively new tool that was being developed throughout the project, so our study served as the "beta test" for the program. Similarly, this study is the first time the All-Sky Survey has been used for research purposes and as such, we carefully examine its utility and efficacy for use in historical research more generally.

## **2. Description of Tools Used**

*The Astrophysics Data System All-Sky Survey (in Aladin):* The ADSASS is a NASA-funded project of Seamless Astronomy aimed at making astronomical data in the form of articles, tables, images, and object references more readily available and more easily accessible to the astrophysical community (Pepe, Goodman, Muench 2011). In conjunction with Seamless Astronomy's mission to study and facilitate the next generation of online astronomical research, the goal of the ADSASS is to link data and literature together through heat maps of article density to further enable "data discovery" (Pepe et al. 2011). The heat maps plot the density of articles for the years 1990-2013 using the astrotags (much like geotags) of the objects mentioned in each article, assigning one astrotag to each object and one count of the article on the heat map for each object. As shown in Figure 1, the densest parts of the heat map are assigned the color red, signifying the astronomical objects or regions of the sky that have the most article counts and therefore are the regions that have been studied the most. As determined by my research,

these red areas could have between about 20 and hundreds of articles. The least dense regions are assigned the color purple, representing about 0 to 10 articles.

**Figure 1**

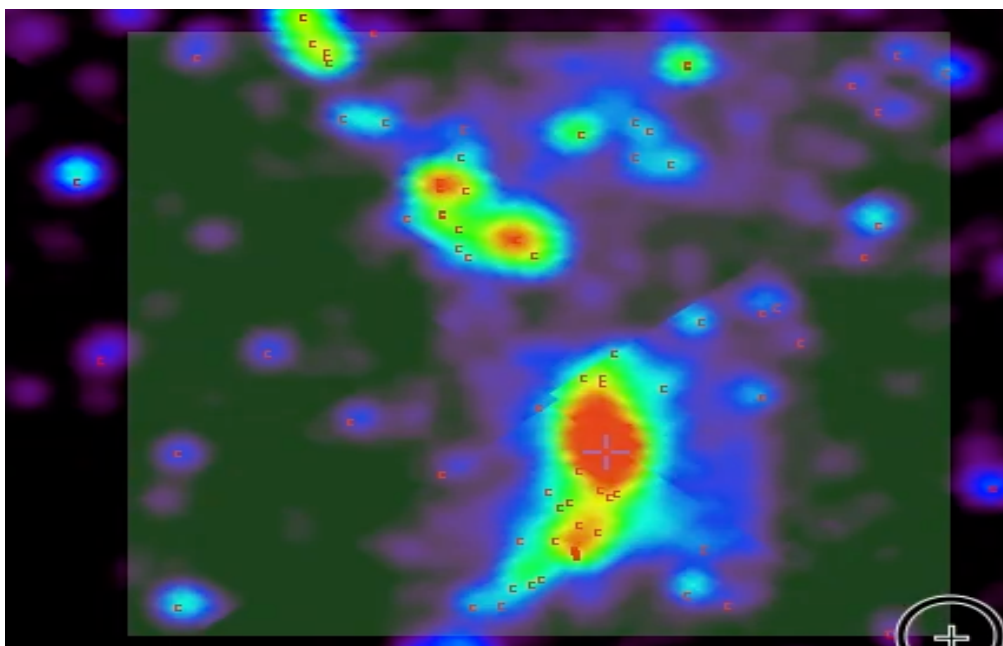


**Fig. 1** shows the Orion Nebula in the ADSASS , in the “All” facet which displays the article density for Orion over all time. The densest regions are those displayed in red, the least dense displayed in purple.

In the ADSASS, the viewer can explore not only a historical heat map by year, but heat maps in other facets by type of object, wavelengths the object was studied in, as well as by base layers, including many surveys at gamma ray through radio wavelengths. The most useful facet for our study was a custom-made layer that represents articles in which at least one author has a Harvard affiliation. Using this facet, we can compare a “Harvard” heat map to the heat map for the rest of the world by toggling back and forth between the two layers. The user can also view the literature associated with any region or object by utilizing the “Select Tool” which, by letting the user draw a rectangular section, shows all objects (as listed in CDS’ SIMBAD service) mentioned in articles as small red squares, demonstrated in Figure 2. After highlighting a region with at least one object in it, a window with a list of all the articles in the region as well as all the objects in that region that are included in the papers will appear (Figure 3). The user then has the option to open all of these articles in ADS in order to read the articles. As of the writing of this

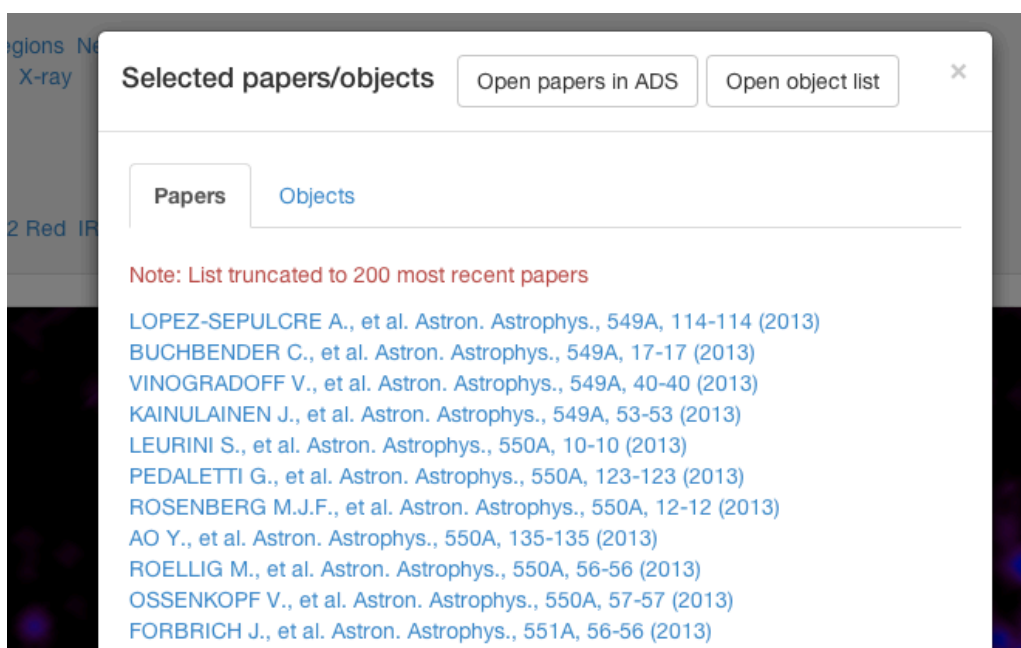
article, the list of articles appears to be truncated, only listing a maximum of 75 articles, which was one of the problems of this study.

**Figure 2**



**Fig. 2** displays an example of use of the “Select Tool.” The mentioned objects appear as small red squares and the user can highlight a region by drawing a rectangular section around the specified area.

**Figure 3**



In **Fig. 3**, the “Select Tool” has been used to produce the literature for all of the mentioned objects in selected rectangular region. The user can also view a list of the objects and can open all of the papers in ADS for further study.

*World Wide Telescope (WWT)*: World Wide Telescope is a computer program designed to allow the user's computer to act as a telescope by combining data from major studies and observations of the sky. Like the ADSASS, the user can toggle between multiple base layers and can view objects and regions of the sky in various sky surveys such as 2MASS. However, WWT provides information about the objects, including distance from Earth, transit and set times, and magnitude, through the "Finder Scope" tool. The user can view individual objects in different survey layers and can then use the "Image Crossfade" tool to adjust the transparency of the survey layer on top of the base layer. Similar to the ADSASS, WWT can open the literature for the selected object in ADS, as well as a list of objects, with the "Research" tool in the "Finder Scope" tool, yet WWT does not yet have direct access to literature heat maps. This study used WWT to identify the objects being analyzed in the ADSASS by inputting the approximate coordinates found in the ADSASS into WWT. We used WWT more importantly to analyze a map of that directly compared the "Harvard" and "All" heat maps. This map, created by Chris Beaumont, consists of a precombined layer of the "Harvard" and "All" heat maps, with the "Harvard" density represented in blue and the "All" density represented in orange, as well as heat maps of the various facets used in the ADSASS. The chosen heat map layer is then put on top of a background layer of the sky and the viewer can change the transparency of the heat map layer.

*The SAO/NASA Astrophysics Data System (ADS)*: The ADS is a digital library for physics and astronomy that contains more than 10.5 million records. Most of this data is bibliographic records and full-text versions of the articles and papers. The ADS aims to make astronomical data research more efficient. This study utilized the ADS for the analysis of papers opened from the ADSASS and to sort these papers by Harvard affiliations.

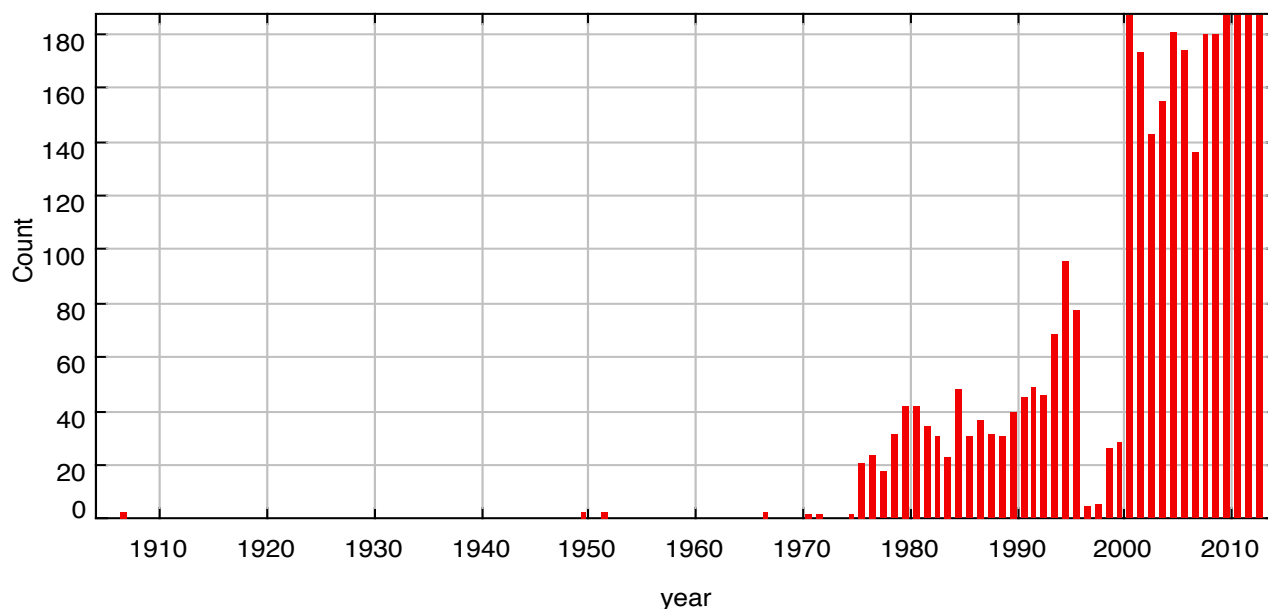
*Centre de Données Astronomiques de Strasbourg (CDS)*: CDS consists of the SIMBAD astronomical database, the VizieR catalogue service of astronomical catalogues and tables, and the Aladin interactive software sky atlas. The ADSASS creates heat maps using CDS' system of astrotagging mentioned objects in the papers.

*TOPCAT*: TOPCAT analyzes and edits large tabular data sets such as source catalogues through an interactive graphical viewer. This study uses TOPCAT to define the limits for our study through determining the completeness of data given to us by Robert Simpson of the number of articles published per year by the CfA.

### **3. Analyzing Processes**

#### **3.1.1 Chosen Limits**

Robert Simpson created a table of 20,000 Harvard-Smithsonian articles, including the articles' bibcodes, names of authors, authors' affiliations, and titles of the articles. We input this table into Topcat and then created a second table including only the unique bibcodes, therefore counting articles with multiple Harvard-Smithsonian affiliated authors only once. Using this new table, we created a histogram (Figure 4) of the number of articles published at the HCO then Harvard-Smithsonian CfA (beginning in 1973) per each year, beginning with the earliest article date of 1895.

**Figure 4**

**Fig. 4** displays a histogram of the count of unique bibcodes of the articles vs. time from 1905 to 2013.

The histogram above in Figure 4 is a clear indicator that our table had incomplete data; we were then able to provide with limits for our study. One of the first signs that the table is an incomplete set of data is the fact that the timeline begins at around 1905 instead of closer to the formation of the HCO in 1839. In addition, the number of articles published in each year before 1975 is far lower than we expected. There is also a large drop in the number of articles at 1997 and a sharp increase at 2000, further leading us to believe this unexpected data demonstrates an incomplete data set. Possible explanations for missing articles could be that authors and scientists in earlier years did not list their affiliation on papers or they had a different or unreadable system for listing affiliations than is used today or the past system is not recognized by ADS. The HCO merged with ADS in 1973, approximately corresponding to the increase in data on the histogram, possibly leading to an influx of scientists, increased funding, and a greater number in projects; therefore, all leading to an increase in papers published. We also cannot include the year 2013



since 2013 has yet to be completed. The most complete set of data appears to be from 2000 to 2012, giving us our limits for the study when using the ADSASS.

To estimate productivity per researcher at the CfA, we sought information about the CfA's (HCO's) total number of scientists over its history. We were extremely surprised to learn from the Human Resources Department and from the Director's Office that this information is not readily available. We were referred to the Computation Facility (CF), who could give us "census" data on what we hoped would be the scientist population, but instead, all the CF provided was the total number of computer users at the CfA, shown in Table 1. From the table, we can see that the drops in the number of scientists in 2001, 2002 and 2006 correspond to drops in the numbers of papers written when compared to Figure 4. This correlation could verify the relationship between the number of scientists present at the CfA and the number of papers published, which would add credence to our choice of limiting our study to the years 2000-2012. Yet, in using the table given to us as displayed in Table 1, we assume that the term "users" does, in fact, mean scientists affiliated with the CfA at that point in time. We also cannot further analyze the census data in terms of names of scientists or with which subgroup in the CfA they are associated.

**Table 1**

Year	Number of Users in Cambridge
1996	745
1997	790
1998	830
1999	880
2000	896
2001	845
2002	840
2003	847
2004	844
2005	861
2006	849
2007	840
2008	886
2009	900
2010	944
2011	921
2012	892

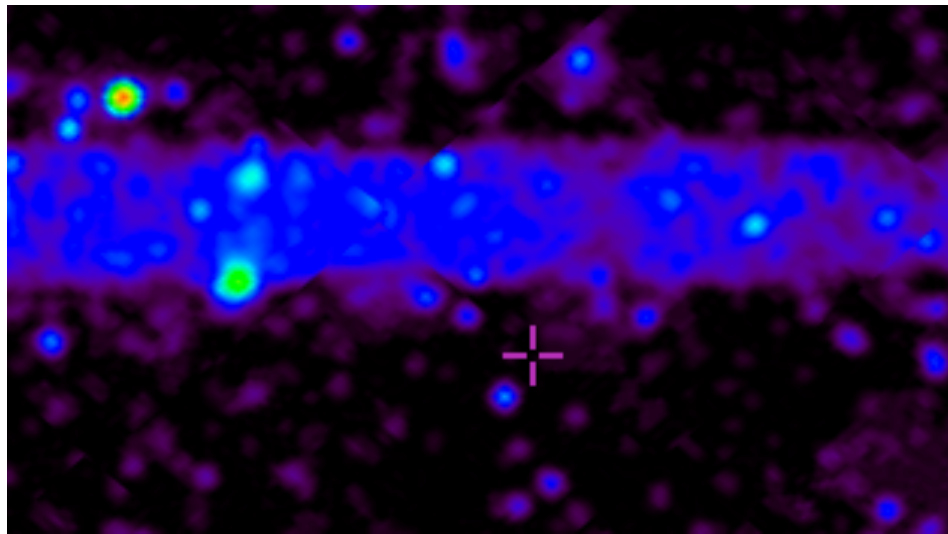
**Table 1** displays the number of users in Cambridge per year, which we assumed to mean number of scientists affiliated with the CfA.

### 3.2.2 Analyzing Processes in the ADSASS

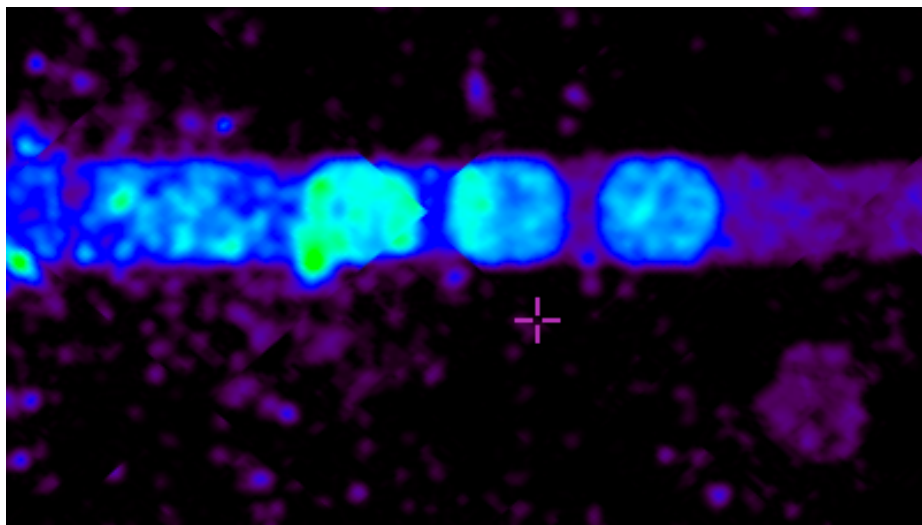
Once the limits were narrowed down, we began the analysis portion of the study. We started the process by identifying the most intensely studied regions on the sky by searching for the “brightest” regions in the Harvard heat map. Once we located a promising region, we viewed the region in the “All” facet to determine if the region is as interesting to the rest of the world as it is to Harvard by comparing the two color schemes of the region. We then analyzed the region in the various facets in order to figure out with what the articles for that region are concerned. For example, one of the densest regions, the Small Magellanic Cloud, is visible in the stars, HII regions, and nebulae object facets and the radio, infrared, UV, XRay, and Halpha wavelength facets. The SMC is also visible in most of the survey base layers, further emphasizing the great extent to which it has been studied. The facets in which the SMC is visible can lead us to determine that it is a region involved with star formation and we can then use the approximate coordinates in WWT or the “Objects” tab in Aladin to conclude that it is the SMC.

In addition to the various facets, we can also analyze the region year by year to see when the object was most studied. The timeline tool is especially useful when attempting to identify specific surveys. Particularly high density in an individual year can signify that the region was a part of a survey. We were able to identify the third largest redshift survey, the 6dF survey, with which Harvard was involved in from 2001 to 2009. Because the results were published in 2009, we saw a sudden high density in 2009, shown in Figure 7, and we then compared this density to the region in the Harvard heat map in Figure 6, which was much denser than the “All” heat map in Figure 5. This discovery led us to determine that this region was a part of some significant research that occurred in 2009 at Harvard. We were then able to confirm that this region was part of the 6dF survey by utilizing the “Select” tool, which listed 6dF as an object and gave us the 2009 published article for 6dF. This process is demonstrated below in Movie 1.

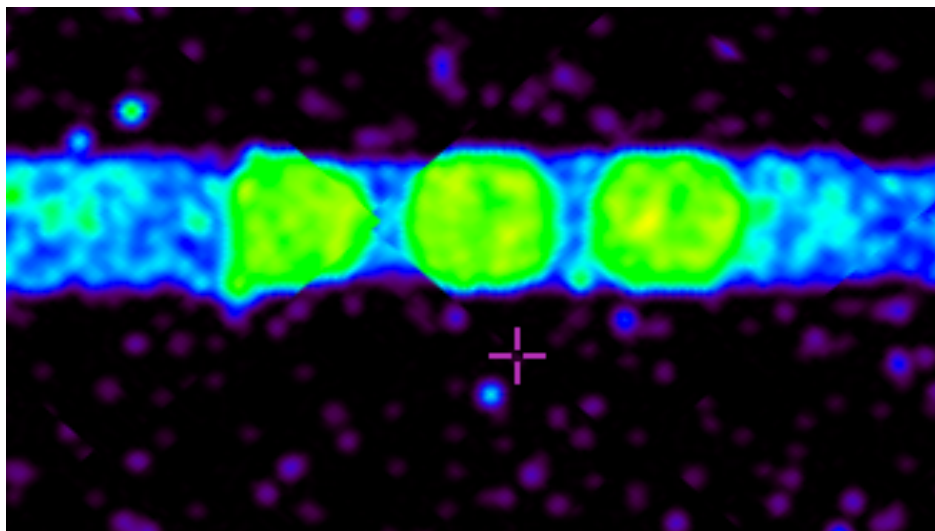
**Figure 5**



**Fig. 5** is a screenshot of the 6dF region in the “All” facet, with the 6dF survey slightly visible as 3 circles in the blue band.

**Figure 6**

**Fig. 6** displays the 6dF survey in the “Harvard” facet, in which the survey is much more visible as circles in the blue band, demonstrating that these circular regions were studied more at Harvard, compared to the rest of the world.

**Figure 7**

**Fig. 7** shows the 6dF region in the year 2009, in which the density suddenly increases, leading us to believe that a study or survey must have occurred at the CfA in that year.



**Movie 1** is a demonstration of how we used the ADSASS to determine that a region is part of the 6df survey that took place at the Harvard-Smithsonian CfA and was published in 2009.

After analyzing a region year by year on the sky, we selected the papers in that region by each individual year from 2000 to 2012 with the select tool and opened the articles in ADS. We recorded the total number of articles for each year, up to the maximum amount of 75 articles shown in ADS, as well as the number of articles with authors who are affiliated with Harvard by filtering the total number of articles by affiliation. If there were Harvard articles for that region for a specific year, we evaluated those articles by the objects included in the articles, to determine if the objects were the same as those with which we classified the region in the ADSASS and WWT. We also evaluated the articles by what wavelengths the studies were

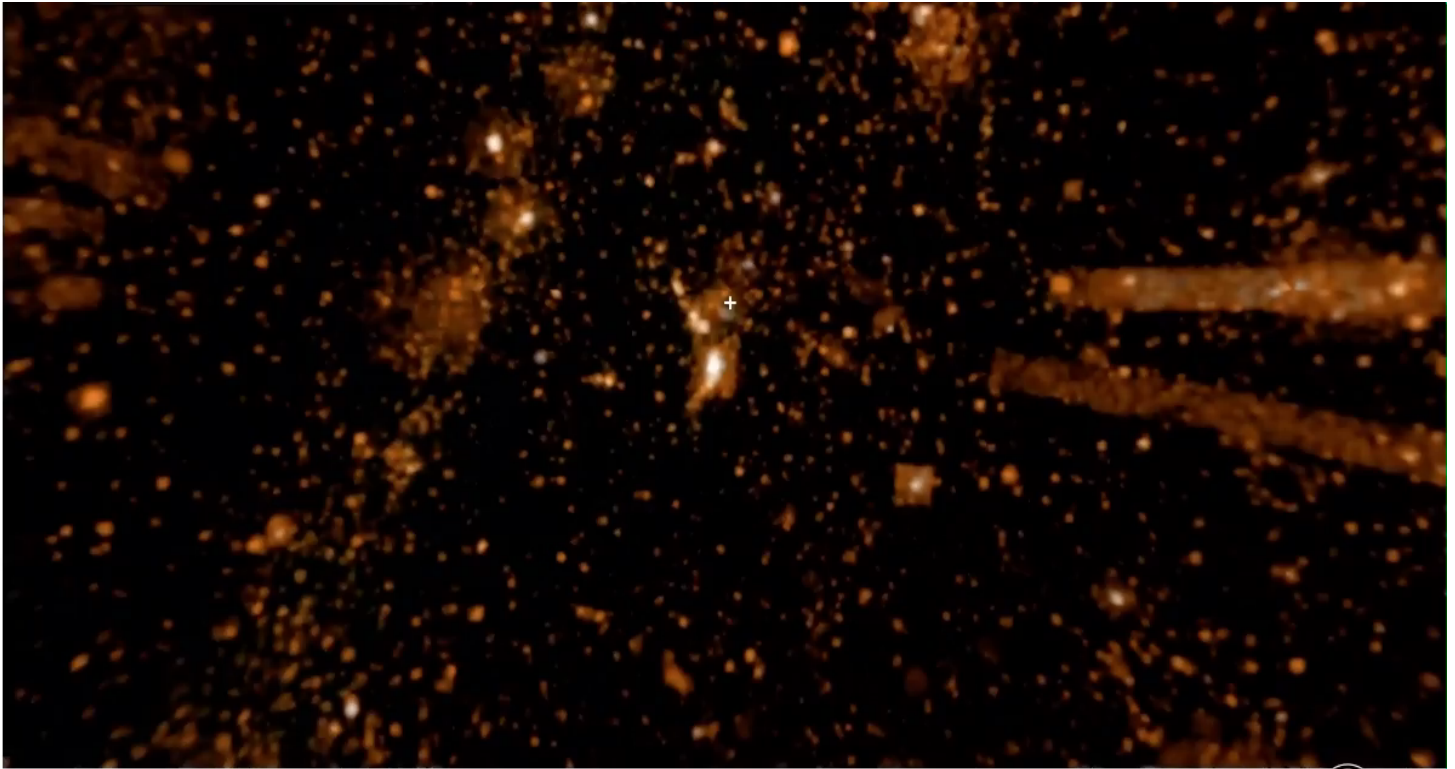
conducted in and if the articles involved any large surveys such as 2MASS or Spitzer. The results of this evaluation process are discussed in Section 5.

#### **4. Observations in Using ADSASS**

The second focus of our project to test the ADSASS led to the identification of multiple improvements for the program as well as the identification of problems in astronomical data science in general. For example, the ADSASS originally did not have any facets for “Affiliation” until Alberto Pepe and Thomas Boch created the custom Harvard heat map using Robert Simpson’s data table, previously mentioned.

The subsequent improvements have resulted in a very usable program suitable for research. At the present, one of the ADSASS’s best features includes the “Select Tool”. We are also now able to toggle between base layers more easily for better comparison between heat maps and multi-wavelength imaging facets. Our end goal is for the toggle tool to be able to overlay transparent layers on top of each other for an even better form of comparison. However, Chris Beaumont created an experimental WWT version which can be opened in WWT and the HTML5-based tool. This version, shown below in Movie 2, blends the heat map for the “All” layer, represented in red, with the heat map for the “Harvard” layer, represented in blue, and overlays this pre-combined layer on top of a base layer in WISE, SFD, IRIS, GLIMPSE, or H-alpha. We were then able to use this version to physically see the differences in density between the “Harvard” and “All” heat maps. This version does not, however, have a “Select Tool”, yet, it is able to be opened in WWT, in which the viewer can use WWT’s Finder Scope.

## Movie 2



In Movie 2, we see a demonstration of the ADSASS in WWT, which utilizes WWT's ability to change the transparency of base layers. We can better compare the "Harvard" and "All" heat maps with a precombined layer contrasting the two heat maps laid on top of a map of the sky.

Even with these greatly helpful improvements, some problems were still encountered in using the ADSASS during research. One of the major issues is not a problem with the ADSASS or WWT in particular but involves the treatment of astronomical objects in research in general. CDS catalogs astronomical objects in articles as point sources (a single RA, Dec position) when assigning the objects astrotags. Instead, many astronomical objects are extended sources that are not well-described and cannot be described by a single "point" coordinate. This problem was apparent in the analysis of the 6dF region, in which the approximate coordinates of the region did not bring up any objects in the Finder Scope in WWT. Yet, this issue is insignificant when studying well-known objects, such as the Orion Nebula, which easily appears in WWT with its

approximate coordinates. Similarly, identifying regions in the ADSASS can continue to be quite difficult as at the moment, there is no way of telling an object studied in only one paper apart from an object studied in many papers. CDS assigns an object an unweighted astrotag and every object mentioned in an article is assigned an article. Therefore, there is no difference between an object merely referenced in an article and an object studied in great length in an article.

Another serious problem for research is that the Select Tool only allows the user to select by individual years and displays a maximum of 200 papers, making accurate comparisons between the number of Harvard papers and the number of papers overall for a region or object difficult. In our evaluation in the “Results” section below, we assume that the Harvard articles produced by ADS are the total Harvard-affiliated article for that region for the specific year and that the maximum limit ADS applies to the number of articles selected has no effect on the number of Harvard papers. We also assume that we selected the same region by hand with the Select Tool for each year and that the area selected remains constant for the entire evaluation of that region.

## 5. Results

We chose to analyze in depth five of the regions we deemed the most dense in the Harvard overlay. We used the ADSASS and WWT to come to the conclusion that these regions include the Large and Small Magellanic Clouds, the Andromeda Galaxy (M31), the W5 Star Formation Region, and the  $h + \chi$  Persei Double Cluster. In evaluating the articles in Table 2, we found that most of the Harvard articles featuring each region had involved studies that are a part of major surveys. Of the 26 Harvard papers written about M31 from 2000-2012, 4 articles were a part of Chandra surveys, another 4 involved the Hubble Space Telescope, and 3 involved 2MASS. M31 is the most studied object out of the regions we analyzed, with 26 total Harvard



articles. The next closest object, the W5 region, is featured in 14 articles; the LMC is featured in 9; the SMC is featured in 8; and the  $h + \chi$  Persei is featured in 5. Numerous articles of these four other objects involve the Spitzer, Chandra, 2MASS and Hubble surveys as well. 6 of the 9 LMC Harvard articles, 2 out of 5  $h + \chi$  Persei articles, 4 out of 8 SMC articles, and at least 3 out of 14 W5 articles involve the mentioned surveys.

Object	Total "Harvard" Affiliated Articles	# of Papers with Chandra	# of Papers with Spitzer	# of Papers with HST	# of Papers with 2MASS	Total # of Papers Including Surveys
M31	26	4	0	5	3	11
W5 Region	14	0	2	0	1	3
LMC	9	0	5	0	1	6
SMC	8	1	0	1	2	3
$h+x$ Persei	5	0	2	0	0	2

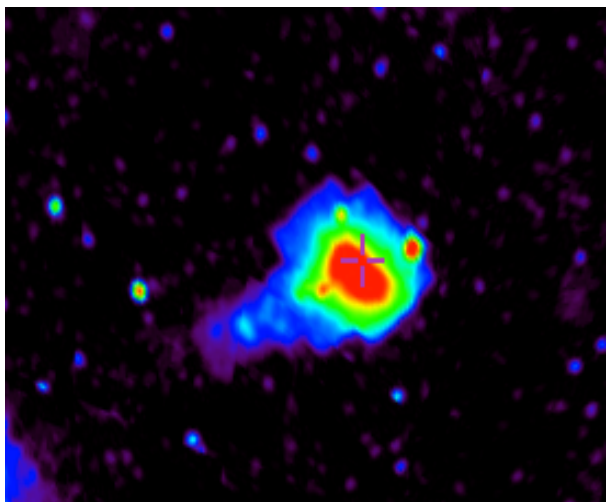
**Table 2** records the analysis done in ADS and shows a trend in Harvard articles involving surveys.

Furthermore, in using Table 2, we discovered that Harvard was the most involved with Spitzer Space Telescope surveys compared to any other survey, with at least 9 of the above Harvard articles heavily featuring studies involving Spitzer surveys. There are most likely more than 9 for these articles, as we expected there to be more W5 articles involving Spitzer than the 2 articles we located, since we know W5 is a star formation region, which are heavily studied in the infrared. Furthermore, we noticed many of the articles featuring other objects in the same regions as the above objects (especially the W5 region) also involve Spitzer surveys.

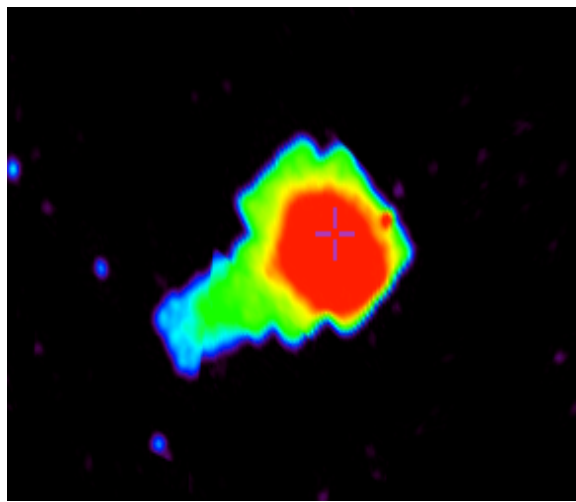
Using the heat maps, Harvard appears to have put more focus on studying the SMC than the rest of the world because the region is more dense for Harvard as shown in Figure 9 than in the “All” facet, shown in Figure 8. We were able to identify the possible origin of this interest for Harvard as the quasars located “behind” the SMC. More Harvard articles focused on these quasars than the actual Cloud, especially in 2003 in which both of the two Harvard articles study

these quasars. This explanation is further supported by the fact that the SMC is very dense in the X-ray facet, as quasars are most often studied in the X-ray wavelength. By the same process, we discovered that one of Harvard's main interests in M31 stems from the study of X-ray binaries. This is also supported by the high density of M31 in the X-ray facet displayed in Figure.

**Figure 8**

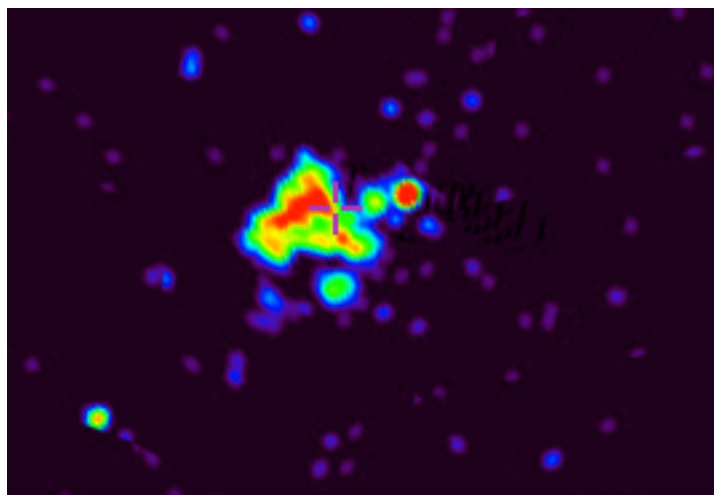


**Figure 9**



**Figures 8 and 9** compare the article density for the SMC in the “All” and “Harvard” heat maps with Fig. 8 displaying the SMC in the “All” heat map and Fig. 9 displaying the SMC in the “Harvard” heat map. We can see the region of the SMC is much more intensely studied by Harvard than the rest of the world.

**Figure 10**

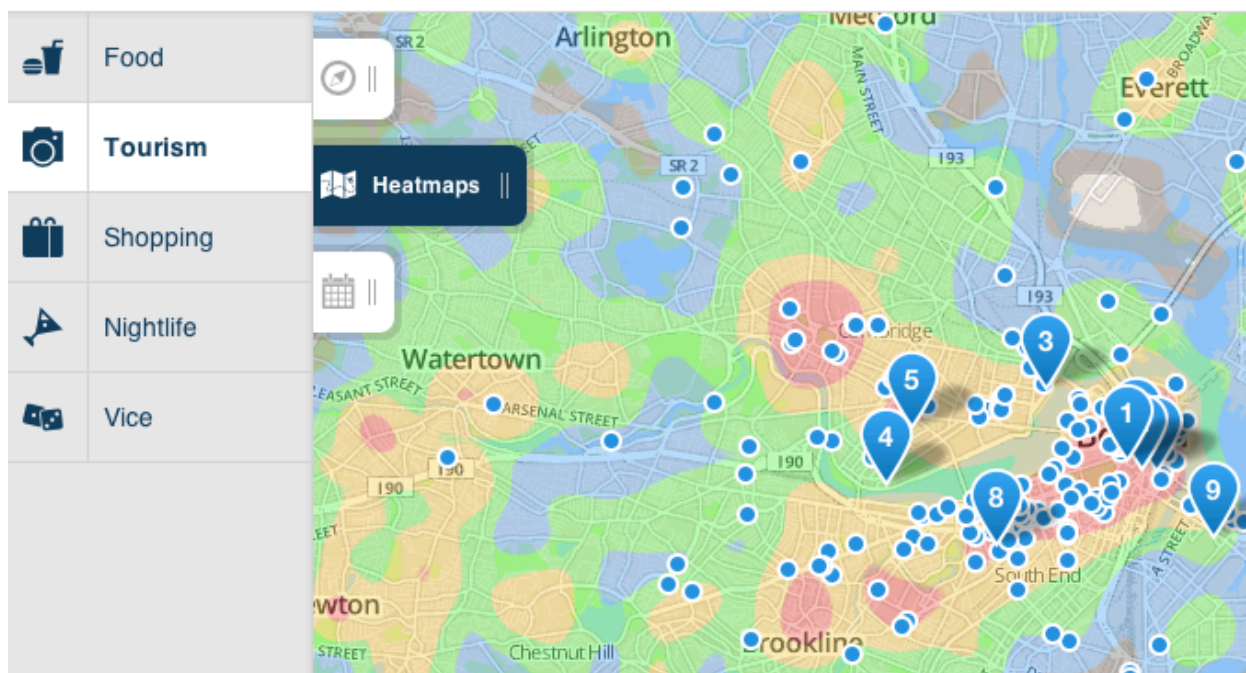


**Fig. 10** displays the SMC in the “XRay” facet and because the SMC still shows a high density in the “XRay” facet, we can conclude that many astrophysicists have studied the quasars behind the SMC region.

## 6. Discussion and Future Projects

This study analyzing the history of the CfA at Harvard University using the ADSASS is the first of its kind to plot literature and data in a heat map. This creates an exciting opportunity to delve into the history of astrophysics in general, but especially for the history of astrophysics at premier institutions, generating publicity for these institutions. As of the writing of this paper, we found few other usages of heat maps in similar ways in other fields. One website, Hipmunk.com, utilizes heat maps to assist customers in travel planning. In Figure 11, the site plots the locations of hotels for a specific city or town and provides heat maps for food, tourism, shopping, nightlife, and entertainment with the most popular areas indicated as peaks in red and the least popular areas indicated as valleys in blue. However, Hipmunk.com does not allow the user to see what exactly comprises the density so the user can see that an area is popular, but cannot see the number of restaurants or shops that make the area dense in the heat map or even the names of those restaurants and shops ([www.hipmunk.com](http://www.hipmunk.com)).

**Figure 11**



**Fig. 11** is a screenshot of a heat map of tourism in Boston and hotels indicated by numbers. The user can identify peaks and valleys in terms of density but does not know what tourism spots these areas actually are.

Another website, Crazyegg.com, is a company that claims to help customers understand their companies' website users by providing heat maps of their customers' websites in terms of where users click. The company also provides "Click-tracking Overlays" that allow the customer to see the percentage of users that click on a region and compare that overlay to an overlay of a separate region, perhaps to determine which area is more popular and more effective. Additional amenities are "Scroll Maps" which show how far users scroll and therefore where their attention fades and "Confetti" which tracks "who clicks what" and user behavior ([www.crazyegg.com](http://www.crazyegg.com)). Crazyegg.com is the most similar program so far found to ADSASS in terms of overlay options and reference point usage. Yet, no other programs seem to combine data, literature, and heat maps on the same scale as ADSASS. This type of program could be very useful in other fields, such as journalism, social studies, and biology.

Future options for continuing this study could be to acquire more specific information regarding census data for the CfA and analyzing the CfA population through its 6 subgroups: Atomic and Molecular Physics; High Energy Astrophysics; Optical and Infrared Astronomy; Radio and Geoastronomy; Solar, Stellar, and Planetary Sciences; and Theoretical Astrophysics. We could use more census data, which we currently do not have due to the ongoing process of digitizing records, to discover more accurate correlations between the number of articles written and the number of scientists at the CfA. It would also be interesting to study how often CfA scientists write papers outside of their subgroups. Eventually, we would like to be able to study the entire history of the HCO and CfA, all the way back to 1839, and use ADSASS for years previous to 1990.

## 7. Conclusion

In using the ADS All-Sky Survey for this study, we were able to identify weaknesses in the program as it is being developed, and thus improve the program for better use for research. The ADSASS is now ready to be utilized for substantial research of the history of the field of astrophysics and for research about specific astronomical institutions in the future. We were able to use the ADSASS to analyze the heat map of the Harvard-Smithsonian Center for Astrophysics and we arrived at several conclusions. In studying 5 of the most dense regions in the Harvard heat map, we found that Harvard has done a significant amount of research using Spitzer Space Telescope Surveys and that many of Harvard's studies involve a survey of some sort, including Chandra and 2MASS. Through this study, tools similar to the ADSASS appear that they would have a large impact on research, especially in combining data and literature.

## References

Pepe, Alberto; Goodman, Alyssa; and Augsut Muench. The ADS All-Sky Survey. Nov 2011