Date submitted: 02/06/2010

Measuring the Impact of the Hubble Space Telescope: open data as a catalyst for science

Jill Lagerstrom
Chief Librarian
Space Telescope Science Institute
Baltimore, MD USA

Meeting: **155. Science and Technology Libraries**

WORLD LIBRARY AND INFORMATION CONGRESS: 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY
10-15 August 2010, Gothenburg, Sweden
http://www.ifla.org/en/ifla76

Abstract

The Space Telescope Science Institute Library and the MAST Data Archive track the productivity (number of refereed papers) and impact (number of citations) of the Hubble Space Telescope in the astronomical literature. I will describe our methodology for collecting and analyzing this data with our HST Bibliography. More importantly, I will illustrate how our publication statistics can be used to show how freely available archival data have shaped the productivity and impact of HST in the astronomical literature steadily over time.

Introduction

In recent years, we have seen much discussion about e-science, open access and science as a public good. For example, the proprietary use of the human genome has created much controversy over public/private ownership of genetic data. (Bentley, 1996). “Big science” and “small science” alike are creating more and more data sets for consumption by researchers (NSF, 2007). The Committee on Data for Science and Technology was formed to “strengthen international science for the benefit of society by promoting improved scientific and technical data management and use.”[1] The field of astronomy provides an excellent model for demonstrating the value of open science and freely available archival data through bibliometric analyses. This paper will describe the work done and the methodology employed at the Space Telescope Science Institute (STScI) by a team of librarians, archive analysts and astronomers for using bibliometrics to show how freely available archival science data have shaped the bibliographic landscape of scientific discovery and data use by the astronomical community. We will show that astronomers are using freely available archival data to perform original scientific work with a high impact in the astronomical literature.

[1] http://www.codata.org/index.html

Astronomers, in general, have been very good at sharing intellectual products. They have been early adopters of many aspects of "open access." For example, their major society journals, which are also the four leading journals in astronomy, *Astronomical Journal, Astrophysical Journal, Monthly Notices of the Royal Astronomical Society* and *Astronomy & Astrophysics* are all freely available electronically 2-3 years after publication. *Astronomical Journal* and *Astrophysical Journal*, both published by the Institute of Physics for the American Astronomical Society, are subsidized by page-charges (the "author-pays model" of open access) which keep subscription prices relatively low. Most astronomers deposit e-prints to the astro-ph section of arxiv.org. The Astrophysics Data System was a pioneer in digitizing historical astronomical literature in the early days of the World Wide Web. This concept of open access extends to astronomical data as well, as we shall see in the following analysis.

To understand the context of our analysis of the impact HST data has made on the astronomical literature, it is first important to understand a few things about the telescope, how observations are selected, and how they are made available to the astronomical community.

The Hubble Space Telescope

The Hubble Space Telescope, now celebrating its 20th birthday, is "a cooperative program of the European Space Agency (ESA) and the National Aeronautics and Space Administration (NASA) to operate a long-lived space-based observatory for the benefit of the international astronomical community."[2] At 17,500 miles per hour at an altitude of 353 miles, it takes Hubble 97 minutes to orbit the Earth. Five space shuttle servicing missions have repaired and upgraded Hubble's suite of instruments, which now include three cameras, two spectrographs, and fine guidance sensors. Hubble observes the ultraviolet, optical, and near infrared wavelengths.

Why put a telescope in space? There are two good reasons: one is to avoid the turbulence of Earth's atmosphere; the other is that infrared and ultraviolet wavelengths are only really observable from space as they are strongly absorbed by Earth's atmosphere. HST has made numerous breakthrough astronomical discoveries and has settled many previously unresolved astronomical problems. The accelerating universe, galaxy mergers, the age of the universe, dark energy, an organic molecule on an exoplanet, and black holes at the centers of galaxies are some of Hubble's top discoveries. 120 gigabytes of science data are downloaded from the telescope every week. To date, HST has produced more than 8,000 refereed papers in the astronomical literature. These papers have garnered more than 300,000 citations.

Data life cycle: from proposal to archive to publication

How does astronomical data from HST get into the hands of astronomers so that they may use observations to make discoveries about the universe and publish these discoveries in papers in astronomical journals? What follows is a simplified

[2] http://www.stsci.edu/hst/HST_overview/

outline of the process. For a complete description of HST observing see the "HST Overview" on the STScI website.[3]

The process begins with a proposal. Unlike some other observatories and unlike many previous NASA space missions, anyone can submit a proposal. There are no restrictions on nationality or academic affiliation. Teams of astronomers submit proposals, which are then evaluated and selected for recommendation for observation by the Time Allocation Committee (TAC). Each HST "cycle" has a TAC dedicated to reviewing its proposals. A cycle consists of approximately one year's worth of observations. A typical twelve month cycle consists of ~3000 orbits of HST around the earth. An orbit produces 50-55 minutes of observation time. Each TAC consists of ~12 panels, each representing a specific science category from solar system objects to star formation to black holes and the far universe, with members from the worldwide astronomical community. The TAC puts the proposals through a rigorous peer review process. In the end, proposals are selected based on a number of criteria including the number of orbits they require and to dedicate observing time to all of the subfields. To give a sense of the highly competitive nature of the process, 958 proposals were submitted for Cycle 17; 228 programs were recommended for approval by the TAC.

Proposals that are accepted for new observations are called "Guest Observer" or "GO" proposals. Up to 10% of the telescope's time is reserved for "Director's Discretionary" time. These programs are mostly devoted to observations of unexpected, important transient phenomena or larger projects of broad community interest. Snapshot programs are survey-type programs whose observations require the duration of less than one orbit. These programs increase the overall efficiency of the observatory because they can be inserted "in between" observations that require a large number of orbits. Finally, parallel programs are observations that use an instrument other than the primary instrument of the running program to collect observations in parallel with the primary science program. As you can see from the above description, every second of the telescope's time is planned in detail to maximize its scientific output.

Once observations are made, data is transmitted from the telescope back to Earth. Staff at the Space Telescope Science Institute in Baltimore process the data for consumption by the proposal teams. Once the data is in the hands of the GOs, they have up to one year of exclusive use of the data to publish papers based on discoveries made with their data. This one year proprietary period may be waived at the request of the proposal team. Director's Discretionary observations are generally made available to the entire astronomical community immediately. Giving proposers a one-year exclusive access to the data creates a balance between the rights of the proposers and the creation of a public good, the HST data archive which is free to all.

Once data are made publicly available, they may be downloaded through the Multimission Archive at STScI (MAST).[4] One of MAST's tasks is to provide data archives and high-level science products to the astronomical community. The primary focus is on optical, ultraviolet and near-infrared parts of the spectrum. Anyone can query the MAST data archive to search for and download HST observations. In

[3] http://www.stsci.edu/hst/HST_overview/
[4] http://archive.stsci.edu/

addition, MAST can be searched with larger meta-search tools such as the NASA/IPAC Extragalactic Database and the Virtual Observatory.

Constructing the HST Bibliography

A metrics team at STScI constructs the HST Bibliography,[5] a searchable database of all HST science publications. The current metrics team consists of the Chief Librarian, Library Technician/Bibliographer, the Archive Sciences Branch Chief, and an Astronomer. To construct the bibliography, we begin by searching the full texts of the major astronomy journals for the name of the telescope as well as its various instruments and surveys[6] using the FUSE Software developed at the European Southern Observatory Library (Erdmann, 2010). After false hits are discarded (HST can also mean "Hawaiian Standard Time") we then evaluate each paper to decide whether or not to include it in the HST Bibliography.

The philosophy of our paper classification is to include those and only those papers that present analysis of HST data to reach a scientific conclusion. In other words, simply quoting results made by others based on HST data does not qualify a paper for HST science paper status. The quantity of HST data involved is not an issue. In fact, sometimes a paper may contain observations from multiple telescopes. As long as HST observations form part of the analysis, we count it as an HST science paper. Additionally, papers that focus solely on the instruments themselves are collected, but not considered part of the science paper bibliography. If a paper is too vague about its data source to be included, we label it as a "grey" paper and do not include it in any of our analyses. We collect both refereed and unrefereed papers, but only do a systematic collection and analyses of refereed titles.

To satisfy our curiosity, we decided, for a certain time period, to also track the number of times HST and its various instruments are simply mentioned in a paper in the four major astronomy journals. This demonstrates HST's broader influence in the literature. In many of these papers, HST science is used to contextualize new research. For example, often a scientist will be motivated by an HST observation to observe a specific astronomical object in a different wavelength with a different telescope.

Once papers are collected, we assign various metadata, some by hand and some programmatically. First of all, we determine the instrument used to collect the data analyzed. Secondly, we identify the program ID from which the data came. Some authors are very forthcoming about the source of their data. Some authors unfortunately are not. For these, we do some behind-the-scenes detective work to figure out which observations they used. Some program IDs prove to be too elusive for us to unearth. This is the case for ~6% of the papers in the database. The bibliographic database is then tied to the proposal database where further analysis can be done on the productivity and impact of the proposals based on number of orbits, science category, etc.

Finally, we determine whether a paper is a "not archival" ("GO"), "partially archival," or "totally archival." These meta-tags help us to assess the influence of

[5] http://archive.stsci.edu/hst/bibliography/

[6] Hubble ACS STIS HST HRS HSP HUDF GHRS GOODS FGS FOS NICMOS FOC WFPC WFPC1 WFPC2 HDF HLA

types of observations in the literature. To do this, we run an automated search that compares the authors of the paper and the authors of the proposal of the data analyzed in the paper. If a paper has no authors that were also on the proposal of the data used, we consider it to be archival. For papers that have authors on both paper and proposal, we consider these to be not archival. Totally archival papers share no authors between the paper and any proposals used by the paper. Partially archival papers use data from some programs that share authors with the paper, but also use data from programs that do not share authors with the paper. Our estimates are conservative: the number of totally archival papers are surely undercounted because there may be cases in which an author uses the data in a new way that was not originally intended when the proposal was executed.

Citation statistics for each paper are imported from the Astrophysics Data System.[7] The ADS is a "Digital Library portal for researchers in Astronomy and Physics, operated by the Smithsonian Astrophysical Observatory (SAO) under a NASA grant." Additionally, we provide information to the ADS so that one may limit ADS search results to the HST bibliography. Moreover, bibliographic records in the ADS link to the MAST archive and provide links to the data used in the paper.

Publication statistics

What can we tell from the statistics generated by our publication database? How are they used to demonstrate the value of the telescope and the impact it has had in the astronomical literature? How do these publication statistics show the value of freely available data archives and how they function as sources of scientific discovery? These questions can be looked at from a number of angles.[8]

To begin with, we can simply show the number of papers and citations generated by the telescope, its various instruments, and programs. Fig.1 shows the number of refereed papers published each year and divides them according to whether they are archival, partially archival, or not archival (GO). A similar plot, which shows citation counts for types of refereed papers, is seen in Fig.2. In addition, we can see the average number of citations per refereed paper for these categories on a yearly basis in Fig.3.

One can see from Fig.1 that in the beginning, right after launch, when the first data were made available, the HST science paper corpus was dominated by GO papers, that is papers published by the team members who won the telescope time that generated the data. As these data become freely available to the community, more and more scientists began to use them in their papers. The number of archival papers has in recent years rivaled those of GO papers. In 2008, there were 303 refereed papers in the archival category and 233 refereed papers in the GO category. More importantly, the impact that archival papers have made closely follows the impact that GO papers have made as is evidenced in Fig.2 and Fig.3. The combined partially and totally archival papers catch up with the GO papers steadily as more data are made available post-launch. Archival papers make up a considerable part of the "market share" in the corpus of HST science papers.

[7] http://www.adsabs.harvard.edu/

[8] For more analyses of HST publications see Apai, 2010.

Richard L. White, Senior Archive Scientist at STScI, has authored a paper (White, 2010) for the forthcoming *Astro2010 Decadal Survey*, a collection of documents written by the astronomical community, which make recommendations to guide the future direction of funding for astronomical research. White teamed up with members of the Chandra X-Ray Observatory, another space mission. He clearly demonstrates the importance and value of the astronomical data archive by using our HST publication database to show the importance of archival papers. His work provides additional analyses, such as examining top-cited papers to show the prevalence of archival work, and demonstrates that archival science is not simply "cleaning up the scraps" in the world of astronomical achievement. The Chandra Science archive, although Chandra is a much younger mission, is showing very similar trends of data usage by the astronomical community. White cogently argues that simply providing free data is not enough; creating user-friendly and accessible archives are what really facilitates science.

Another way to illustrate the widespread use of HST data is to look at the affiliations listed in HST science articles. Table 1 shows the breakdown by country of the affiliations listed in totally archival HST science papers from 2008. In the 303 papers for that year, 1714 affiliations are listed.[9] Affiliation data in the ADS are not complete. Regardless, we still feel that generating a list of affiliations, however incomplete, shows the breadth of use of HST data. 38 countries, from each inhabited continent, can claim credit for papers written with totally archival HST data in 2008.

What kinds of discoveries can be made with archival data? In 2009, a team of astronomers at the University of Toronto led by David Lafreniere, developed new techniques to process astronomical data and, as a result, uncovered a hidden planet in an archival Hubble image.[10] Who knows what other planets are lurking in the archives? New and innovative data processing techniques may lead to further discoveries in "old data." HST archival data has been put to good use closer to home as well. In 1998, a team led by Robin Evans of JPL in California used archival images to find new asteroids.[11] These were completely serendipitous discoveries --- asteroids were not the primary science targets of the observations. 96 objects were reported to the International Astronomical Union's Minor Planet Center. Of these, most were newly discovered asteroids; new data about known asteroids were used to update information about their orbits. Our last example involves a supernova.[12] Supernovae progenitors are very rare discoveries. Astronomers have theorized what a star looks like before it explodes, but have rarely been able to find the data to analyze these progenitors. In 2005, a star exploded. Thankfully, HST had taken pictures of the star in 1997 for different scientific reasons. The archival data was able to provide a true "missing link" in astronomical research.

[9] Note that this is a total listing of affiliations, not distinct authors or papers. Our data shows that in this sample Lithuania was listed as an affiliation four times. This could be for one paper or four, depending on the combination of co-authors.
[10] http://hubblesite.org/newscenter/archive/releases/2009/15/full/
[11] http://hubblesite.org/newscenter/archive/releases/1998/10/
[12] http://hubblesite.org/newscenter/archive/releases/2009/13/

Conclusions

With our carefully constructed bibliographic database, we can clearly demonstrate the value of archival papers through bibliometric analyses. The productivity and impact of archival papers rival that of papers authored by those who have exclusive use of new data for a one-year period. Authors from 38 countries around the globe have made use of non-proprietary data in 2008 alone. We can also qualitatively show, through example, the serendipitous use of archival images and that new and innovated techniques can be used to make new discoveries with "old" data.

One may ask – are observatories with archives more productive and do they provide a greater impact? According to Virginia Trimble's latest research (Trimble, 2010), which uses fractional counting to assess productivity and impact in major astronomical journals, papers which re-used only archival observations in 2008 had a slightly higher citation rate. Comparing telescope bibliographies must be done carefully because different observatories have used varying methods and criteria for creating publication lists.[13] Emphasizing this point, a recent paper in *Nature* (Lane, 2010) calls for "making science metrics more scientific. " Moreover, a recent study by the author (Lagerstrom, 2010) has revealed issues that need to be resolved in the reliability and validity of telescope bibliography data sets if we are to fairly compare. Thankfully, efforts are now being made to establish some standards and best practices for creating these bibliographies. (For further information see Grothkopf, 2010 and Kitt, 2010).

The title of this conference is "Open access to knowledge - promoting sustainable progress." It should be evident from the description and analyses provided above that astronomers and librarians together value the importance of sharing knowledge openly and freely. Our next challenge will be to ensure the preservation of this data for future generations. Moreover, it is important that we do more to share and preserve individual astronomers processed data sets in addition to the "raw" data we currently share openly. It is hoped that the astronomical community can serve as a model and inspiration for those who wish to promote open sharing of the products of science and that our HST Bibliography can be used to take steps, or even giant leaps, in this direction.

Acknowledgments

STScI metrics team : Daniel Apai, Jill Lagerstrom, Elizabeth Fraser, Karen Levay and Alexis Truitt. Thanks to the HST Mission Office for travel funding to the 2010 IFLA conference. And to Rick White for the Astro 2010 paper which shows the impact of HST's data archive. This research has made use of NASA's Astrophysics Data System.

[13] The European Southern Observatory Library lists various statistics gathered by observatories on its website:
http://www.eso.org/sci/libraries/edocs/ESO/ESOstats.pdf

References

Apai, Daniel; Lagerstrom, Jill; Reid, Ian Neill; Levay, Karen L.; Fraser, Elizabeth; Nota, Antonella and Henneken, Edwin. (2010) Lessons from a High-Impact Observatory : The Hubble Space Telescope's Science Productivity between 1998 and 2008. Publications of the Astronomical Society of the Pacific, in press.

Bentley, David R. (1996) Genomic Sequence Information Should Be Released Immediately and Freely in the Public Domain. Science 274, 533-534. Accessed at : http://www.jstor.org/stable/2899624.

Erdmann, Christopher and Grothkopf, Uta. (2010) Next generation bibliometrics and the evolution of the ESO Telescope Bibliography, in: Library and Information Services in Astronomy VI, Isaksson, E., Lagerstrom, J., Bawdekar, N. and Holl,A. (eds.), Astronomical Society of the Pacific, San Francisco, ASP Conference Series, in press.

Grothkopf, Uta and Lagerstrom, Jill. (2010)Telescope Bibliometrics 101, in : Future Professional Communication in Astronomy II 13-14 April 2010, forthcoming.

Kitt, Sandra and Grothkopf, Uta. (2010) Telescope Bibliography Cookbook: Creating a Database of Scientific Papers that Use Observational Data in: Library and Information Services in Astronomy VI, Isaksson,E., Lagerstrom, J., Bawdekar, N. and Holl, A. (eds.), Astronomical Society of the Pacific, San Francisco, ASP Conference Series, in press.

Lagerstrom, Jill. (2010) Comparison of methods for creating telescope bibliographies, in: Library and Information Services in Astronomy VI, Isaksson, E., Lagerstrom, J., Bawdekar, N. and Holl, A. (eds.), Astronomical Society of the Pacific, San Francisco, ASP Conference Series, in press.

Lane, Julia. (2010) Let's make science metrics more scientific. Nature 464, 488-489. doi:10.1038/464488a.

[NSF] National Science Foundation. (2007)Cyberinfrastructure Vision for 21st Century Discovery. Access at : www.nsf.gov/pubs/2007/nsf0728/index.jsp.

Trimble, Virginia and Ceja, J.A. (2010) Productivity and impact of astronomical facilities : a recent sample. Astronomische Nachrichten 331**,** 338-345, doi: 10.1002/asna.200911339.

White, Richard L.; Accomazzi, Alberto; Berriman, G. Bruce; Fabbiano, Giuseppina; Madore, Barry F.; Mazzarella, Joseph M.; Rots, Arnold; Smale, Alan P.; Storrie, Lombardi, Lisa; and Winkelman, Sherry. (2010) The High Impact of Astronomical Data Archives in: Astro2010: The Astronomy and Astrophysics Decadal Survey, Position Papers, no. 64, in press.

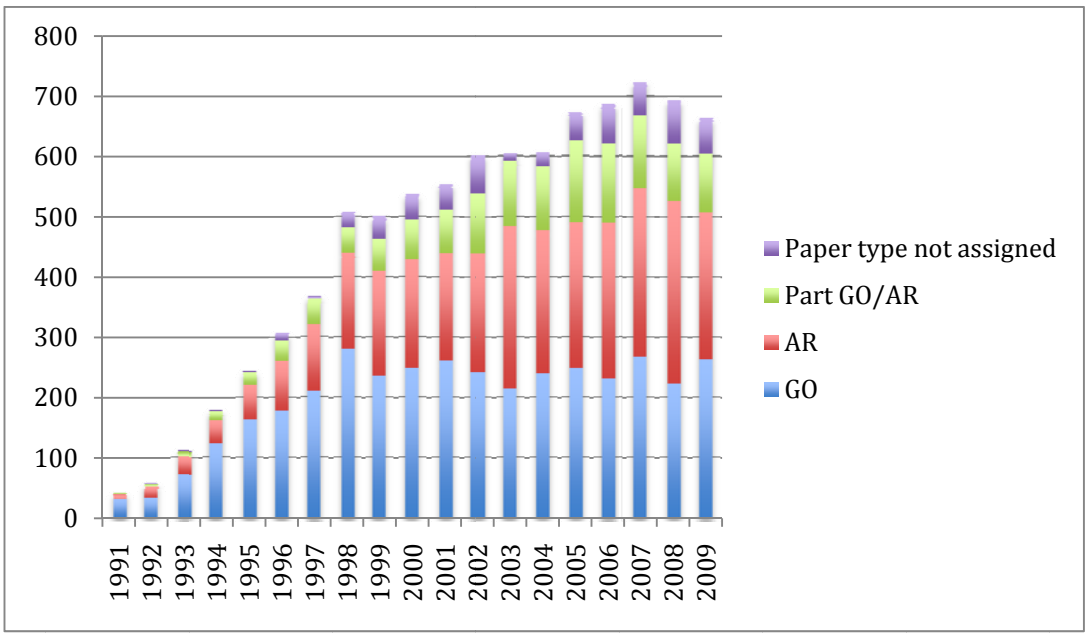

Figure 1. Number of papers per year by paper type.

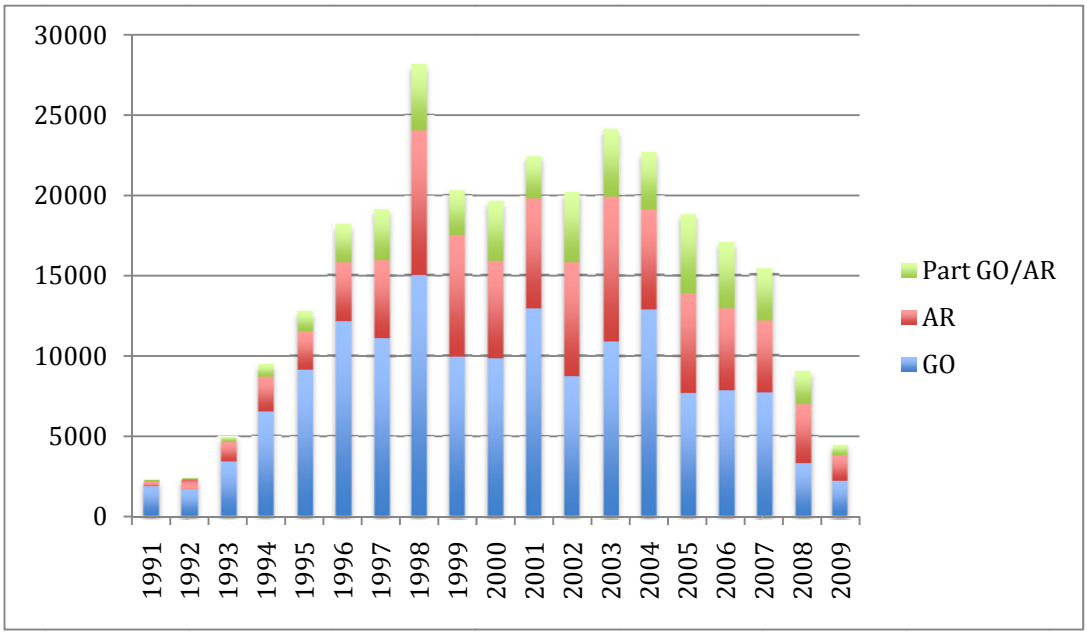

Figure 2. Number of citations per year by paper type.

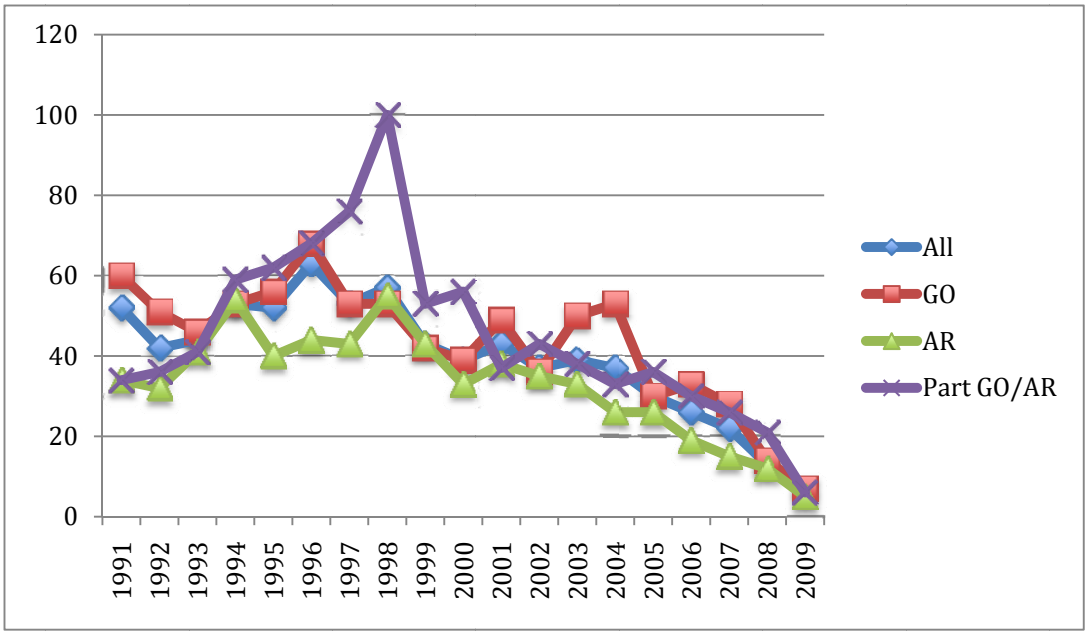

Figure 3. Average citations per paper per year by paper type.

Country	Number of papers	Country	Number of papers
Cyprus	1	Finland	17
New Zealand	1	Russia	17
Portugal	1	Brazil	18
Singapore	1	Belgium	20
Ukraine	1	Denmark	21
Hungary	3	Sweden	23
Georgia	4	Japan	30
Israel	4	Korea	33
Lithuania	4	Chile	38
Austria	5	Canada	39
Ireland	6	Australia	40
Poland	6	Mexico	42
Switzerland	6	Netherlands	44
Greece	7	Spain	71
Bulgaria	8	UK	112
Argentina	9	France	121
Taiwan	10	Germany	174
India	11	Italy	196
South Africa	14	US	529

Table 1. Number of country affiliations listed in 2008 totally archival papers

Hubble Space Telescope. Final Release Over Earth after Servicing Mission in 2009. Credit : STScI

HST 20th Anniversary Image Release. A mountain of gas and dust rising in the Carina Nebula. Credit: NASA, ESA, and M. Livio and the Hubble 20th Anniversary Team (STScI)