Handling, archiving, and citing data in astronomy

Alberto Pepe, Alyssa Goodman, August Muench, Merce Crosas, Christopher Erdmann

1 Abstract

We report the results of interviews with astronomers at the Harvard-Smithsonian Center for Astrophysics. [?] [?]

> No, I don't have a website where I store these data. Most of it is in various stages of mess.
> —An Astronomer

2 Introduction

Astronomers produce and peruse vast amounts of scientific data. Making these data publicly available is important to enable both reproducible research and long term data curation and preservation (King, 1995, "Replication, Replication", Political Science and Politics, 28: 443-449). Because of their sheer size, however, astronomical data are often left out entirely from scientific publications and are thus hard to find and obtain. In recent years, more and more astronomers are choosing to store and make available their data on institutional repositories, personal websites and data digital libraries.

Just to show how citations work, here is a cite to Batista's work [?] and Leo Egghe's [?]. While here is a citation which is not even in the bibliography file but it is on ADS so it can be cited by URL [?].

we describe the use of personal data repositories as a means to enable the publication of data by individual astronomy researchers. by repository we mean

in astronomy this accumulation might include the collection of bits of raw images taken at the telescope or subsets of processed data from a space observatory archive.

from this collection or pile of data, the data stack is distilled into new research objects. for example, raw spectra are calibrated and combined into a higher s/n data product.

these distilled products are further refined even chopped up into smaller bits where the relevant scientific information packet is much more highly concentrated; we consider such a packet of knowledge publishable

consider this flow of information then consider just how linear it appears to be.

the typical end of this evolution of accumulation and distillation the research data is the publication.

there are a few problems with data objects appearing in papers: if at all they capture the most refined research objects. they fork only with the paper. the avoid curation by domain specific experts – the journals have neither a peer-review process nor an editorial process for data.

worse, they are not trackable in the papers. Even if they do have identifiers and even if those identifiers , these data products require a different framework for reuse.

By data materials, we mean any data product available on the web which was either instrumental for the pursuit of research, e.g. raw data from astronomical archives, or generated in the context of research, e.g., reduced and processed data presented in a paper.

3 Results

3.1 Exploratory analysis of data citation practices

To begin, we mine a corpus of astronomy articles for external web links. By "external web link" we mean: any outgoing link embedded in the final published version of an article (e.g., its PDF or HTML format) which points to an online resource in the `http` (or `https`) URI scheme. The purpose of this exploratory analysis is to assess whether astronomers use links within articles to point to datasets and related supplemental data resources.

We analyze a corpus of all articles published in the four main astronomy journals (The Astrophysical Journal, The Astrophysical Journal Letters, The Astrophysical Journal Supplement, The Astronomical Journal) between 1997 and 2008. We find a total of 13447 potential links to datasets in a total of 7641 publications. The detailed procedure by which potential data links are selected and filtered is described in the Materials and Methods section.

In the barplot of Figure **??** we show how linking practices have changed over time. Links to potential data resources in astronomy first appear in 1997, with only a couple of dozens links published in that year, and quickly increases every year to reach around 1500 yearly links in 2005. After 2005, the volume of total published links roughly stays the same every year. The graph shows that with widespread use and adoption of the WWW, linking to online resources within published articles has become more and more popular. The bars in the barplot of Figure **??** also depict whether published links are still available as of December 2011: the green portion of each bar represents the volume of valid links (HTTP status code 200: OK), while the grey portion of the bars represents broken links (HTTP status codes 3xx, 4xx, and 5xx). This link categorization shows that half or more of all links published prior to 2001 are now broken. The percentage of broken links decreases with time to reach roughly 10% in 2008: one in ten links included in astronomy papers in 2008 is unreachable three years later.

This analysis can be pushed further by exploring two distinct subsets of the astronomy link corpus. In Figure **??** we show how the percentages of broken links differ over time for a set of 1801 links to personal websites (links which contain the tilde symbol ~ which are usually reserved for personal web pages on institutional servers) and a set of 3731 links to institutional, curated archives (a manually selected list of domains that are obvious astronomy archives, such as `archive.stsci.edu`). Attempting to make a distinction between these two categories of links is of crucial importance. The former set of links, the "tilde links", are potential pointers to datasets found on personal websites. These may consist of data tables and images which are the product of data analysis and reduction procedures described in the accompanying paper. As such, they do not belong to larger curated archives, which normally host raw data only. Ideally, these datasets would be included in the full text of the article, but oftentimes they are too large to fit within the format of a published paper and are included on a personal server and linked from within the paper. The latter set of links, the "curated archives" links is, instead, a collection of pointers to established archives and repositories, managed and curated by institutions, surveys, telescope sites. Authors may want to link to these resources to cite and acknowledge the raw data sources that they employed in their research. Figure **??** shows that the availability of these two categories of links follow very different, yet expected, patterns. The vast majority of "tilde links" published between 1997 and 2003 is not available any more (personal links are depicted as a black solid line and circles). Astronomers change locations, jobs, institutions and, as such, their personal web servers change or expire over time. However, the percentage of broken links to personal websites falls rapidly: nearly all "tilde links" published in 2008 are still accessible today. A different scenario emerges when one looks at the temporal pattern for links to curated archives (depicted in the graph as a red line and crosses): the percentage of broken links stays roughly the same over time (between 15% and 20%), indicating that curated, institutional websites are much less vulnerable to temporal effects than personal websites.

This exploratory analysis reveals three key findings. First, since the inception of the web in the early 1990's, astronomers have increasingly used links in articles to cite datasets and other resources which

do not fit in the traditional referencing schemes for bibliographic materials. Second, as for nearly every resource on the web, availability of linked material decays with time: old links to astronomical materials are more likely to be broken than more recent ones. Third, links to "personal datasets", i.e., links to potential data hosted on astronomers' personal websites, become unreachable much faster than links to curated "institutional datasets".

These findings point to a preliminary realization: that astronomers do have a need to reference and include data materials in their published work. Since they lack a standardized mechanism to reference these resources — data citations do not normally fit in the format, structure, and scope of published journal articles — they attempt to cite datasets using simple linking from within articles. Results from this preliminary analysis prompted a qualitative interview study, described below.

3.2 Interview results

We conducted interviews with a dozen astronomers of the Harvard-Smithsonian Center for Astrophysics working in disparate fields of astronomy and at different stages of their career: postdoctoral researchers, staff scientists, tenure-track and tenured faculty. All interviews were conducted in person between March and July, 2011.

The purpose of the interviews was to gather a first-hand account of the needs and challenges of data referencing and archiving in astronomy. Our interview rubric was freely based on the Data Curation Profiles Toolkit developed by the Distributed Data Curation Center at the Purdue University Libraries and the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign (`http://datacurationprofiles.org/`). Before every interview we created a record of the interviewee which contained key information such as name, academic role, affiliation, department, area of specialization, website, as well as an annotated list of recent and/or prominent astronomy projects pursued and published datasets, and pointers to one or two recent published articles, possibly containing links to daatsets. The template for our semi-structured interview consists of questions revolving around these topics:

A story We begin with a very open-ended question, asking astronomers to tell us a story about their data. In the case of very prolific authors, we ask them to focus their story around a specific paper or project. We allow the researcher to discuss about their research, their data practices, their data output, their scientific work flow, and their community of practice. With this first question, we gauge potential projects and paper and we steer the conversation towards a specific one, which becomes the subject of the following questions.

Generated output What were the important stages of data production, analysis and interpretation? Did you collect new data? Archival data? How dependent are your results on the software tools used in each stage of the data analysis? Did you create new software?

Availability Are any/all of these data currently available for download/perusal? If yes, where? What platform are you using? What stages, versions or types of the data are available? If not, why not? Would you be happy to make those data available?

Data citation How can your data be cited/referenced? Can you pinpoint some publications that were clearly based on these data? Are these publications on ADS?

Format and size Are the data available as separate files? What formats are they in? How large are they?

Ownership What sort of licensing do you envision for your data? Do you have contractual obligations and/or restrictions to preserve or share your data?

Desired features If your data were to be made available on a platform that allows their storage, discovery, and citation, would you want to offer visualizations of your data? Would you want to allow users to run simple statistical analyses on your data? Would you allow users to download the entire datasets or portions of thereof?

3.2.1 Data stories

During the interviews, we listened to a very diverse collection of data stories. In most cases, the stories were very much rooted not only in the specific project that we were being told about, but in the data practices of a given subdiscipline of astronomy. For example, an interviewee working with quasars monitors and regularly publishes flux density data which are used for calibration purposes. These data are relatively limited in size and are hosted on an institutional webserver:

> There is a website which is essentially a flat ASCII file that has information for a particular day for a given number of quasars. I convert the raw data into a standard format with columns: source, date, time frequency, flux and error.

Another example is an interviewee working with galaxy clusters who told us that the amount of data handled and processed in their research is so large that it involves the joint work of many staff scientists and graduate students. Hosting and providing access to the various levels of data involved in the production of the final reduced data is beyond the capabilities of a single research group. In their own words:

> We could certainly put a data table in the publication with very heavily digested quantities like velocity dispersion and number of galaxies, but those things are derived from upstream raw data. You would argue that it would be more value to the community if we were to make the image archive available. I am probably not going to send all the Magellan and HSST images to the ApJ though. But I could well imagine twenty years in the future that that image archive has more endured value than our attempt to extract information out of those images.

These two examples are telling of the differing scales at which data practices operate: from small continually-updated datasets which are currently hosted on personal webservers to large, collaboration-enabled surveys whose data do not have an obvious home. Overall, we found that the mechanisms by which data are used and handled differ widely from project to project and between different subdomains and wavelengths.

3.2.2 Generated output

As for the previous question, the data products generated in the context of different research endeavors, and their prodcution mechanisms, varied greatly between different projects. An interviewee, for example, indicated that the source of their research is entirely archival data and that the bulk of their research is writing the software and running analyses with it:

> We just used and combined catalog data from many different large area surveys containing photmetric description of different extragalactic sources (galaxies and quasars): their magnitude, fluxes, and morphological parameters. Then we subjected these large tables to some Machine Learning methods to estimate the redshift of the sources. The result was an augmented table which included additional information about estimates of photometric redshits.

In some other cases, astronomers were interested exclusively in the scientific findings of their research; the mechanisms by which the data were reduced and analyzed might have not been documented properly:

> We didnt write software from scratch, but we used it in ways that might not be so easily reproduced. Thats what you read in the data section of a paper when it says something like: *we smoothed the data to such and such a resolution and then we did this and then we did that.* Whether the person [running the analysis] gets the order of the steps right may actually affect the final outcome. I am not sure whether these software workflows got perfectly documented.

Despite the many types of data products generated, a visible thread of similarity between responses can be found in the prominence of social and human factors involved in the production of these data products. Interviewees often reported that the various levels of data generated are entirely in the hands of the people involved in the projects. An interviewee summarized the prevalence of this practice as:

> If we were rich and organized we would be like Sloan and we would have: Data release 1.0, Data release 2.0, etc. But we have more like: Graduate student 1, Graduate student 2, Graduate student 3 (laughs)

3.2.3 Availability

All the astronomers interviewed in this study state that they are willing to share with the public all the reduced data generated in the context of the discussed projects. Only two-thirds of them, however, have gone through the effort of storing the data and making it available online.

The vast majority of those that currently make available their reduced data online chooses to use a dedicated personal webserver, generally accessible from the Principal Investigator's personal website or group laboratory page. The flavors and levels of data offered on these personal webservers differs greatly among projects. however. Some astronomers limit themselves to posting the minimum amount of data necessary to supplement a published article, or to accommodate the requests of the referees to see the data. In some other cases, astronomers post various levels of data, from raw to reduced data. Yet, whether the amount and description of data suplied is sufficient to entirely replicate a study is unclear and varies from case to case. One astronomer admits that access to raw data is a barrier to reproducibility of results:

> Could we get the raw data from that survey? We did not archive the totally raw unreduced data but there is a tape library somewhere with all the data, but it would be difficult to find. And so Id give you maybe sixty percent odds that we could get that data now. Those raw data were taken in 2001, 2003, 2004, and maybe some in 2005. I dont even remember.

Another astronomer working with raw data from a larger survey (Sloan Digital Sky Survey) indicated that the raw data used in their study are indeed available somewhere (on the SDSS archives), but has doubts on whether linking raw to reduced data has a real utility:

> How many people re-reduce SDSS images? I make a guess: there are probably ten people on the face of earth that ever re-reduced Sloan images.

Only a couple of interviewed astronomers employed other techniques to make the data available, which do not involve posting data to a private webserver. For example, the catalogs of photometric redshifts discussed earlier on were made available via dedicated services in the VO framework (Virtual Observatory). They can be accessed through the VO registry and through a number of popular astronomy applications.

3.2.4 Data citation

Interviewees are also unsure about the best way that other researchers can cite their data. If they have published a "data paper", i.e. a refereed article describing the data, the data collection, and analysis in

detail, they prefer to receive a citation to the paper. In all the other cases, they are happy to just receive mention of the via a URL link pointing to the data or an acknoweldgement in the publication.

> Journals dont seem to be concerned with standardizing that [how data are cited]. If you use the data from someone elses project then we just say we downloaded it from the archive. Sometimes people cite the program number and other times people go through the trouble of seeing if a paper has been published on it.

3.2.5 Format and size

All astronomers unanimously indicated FITS (Flexible Image Transport System) to be the data format of choice for all their data needs. As one astronomer aptly summarized:

> The FITS format does everything I need. It's hard to change. It is a ubquitous self-defining data structure. You can download one from 20 years ago and it still works.

As for size, the spectrum was much more diversified with some small datasets, e.g., in the range of few Megabytes for quasar density flux data, some medium-sized datasets, e.g., up to a dozen Gigabytes total for the thermal emission data from the survey of star forming regions, to some much larger archives in the order of many Terabytes, e.g., for galaxy cluster image data.

3.2.6 Ownership

Astronomy is a discipline which studies a matter — celestial objects and astronomical phenomena — that are by definition public domain. This is probably why the inclination to share data seems to be ingrained in the mindframe of virtually all astronomers.

None of the interviewed researchers indicated that the data were "theirs" or that they were under contractual agreements of working under restrictions that would impede them to share their reduced data. All astronomers indicated that their data, no matter how reduced and ingested from its original raw format, were public data. This remark was stressed even more by two interviewed "computational astronomers" whose research is based on the aggregation and analysis of data in existing astronomical catalogs:

> We truly believe that sharing data is the right thing to do, simply because the original data we used for this study was not ours. Our study was only possible because other astronomers made their data publicly available in the first place!

3.2.7 Desired features

We asked astronomers whether they could think of any specific features that an online hosting platform for their reduced data should have in order to allow easy access, visualization, and analysis by users.

All respondents indicated that such a platform should, at the most basic level, allow citation and download of the data. Another very basic feature suggested by nearly every interviewee is the ability to select and download only a subset of the data available for a specific project, rather than the entire dataset. Thus, for example, a user should be able to select a region of the sky delimited by coordinates (Right Ascension, Declination and an angular radius) and download matching observations for that region. For time-varying phenomena, the ability to subset by temporal parameters was indicated. Only a small portion of the people interviewed indicated the need for a more sophisticated filtering and subsetting mechanism, supported by a strong query language and/or full interoperability with existing frameworks, such as the VO registry.

Interestingly, none of the interviewed astronomers suggested that the data hosting platform features advanced analysis and visualization techniques.

4 Discussion

We find that astronomers are increasingly willing to reference and share the secondary or processed data sets used to derive the results in their publications. However, a common infrastructure to share this type of data sets and guidelines for good practices on how to cite them are still lacking. This results in invalid data references over time and incomplete publications which can not be validated or built upon them.

This group is involved in a project that has provided a solution to these problems in social science (refs), and is now in the process of being adapted to astronomy (theastrodata.org, seamless astronomy refs). The project, which uses the Dataverse Network software as the underlying infrastructure (refs), intends to achieve two main goals, both critical in data sharing;

1) a central repository where (small) astronomy data sets can be deposited and archived for long term access, and 2) a data citation that includes a persistent identifier which links to the data, and should be added to the the references sections of any publication.

The central repository not only serves as a mere file system to drop and access data files, but instead provides the tools to understand the nature of the data sets and how they can be reused. It accomplishes this by allowing to add descriptive metadata about the data set and complementary files such as documentation and code, and extracting metdata automatically from the data file. It also provides the infrastructure to replicate the data files to multiple locations and export the metadata to make the data sets more easily discoverable by other systems.

A formal data citation is the other key piece of data sharing. It provides a persistent link between the publication and the data set, so that if the location of the data set changes in the future, the persistent link can still be resolved to the same data set (ref. to Handles). It also provides attribution to the various contributors - authors and data producers or providers - properly given credit to the authors that collected and process the data. Finally, a formal, standardized data citation is needed to facilitate the adoption of data citation by publishers - it is critical that this type of citations become part of the references sections in publications, and are easily traceable to derive their impact.

5 Materials and Methods

We analyze a corpus of all articles published between 1997 and 2008 in the four main astronomy journals (The Astrophysical Journal, The Astrophysical Journal Letters, The Astrophysical Journal Supplement, The Astronomical Journal) which contain external URL links in their full text. We initially find 33847 external links in 13390 articles. `http://hdl.handle.net/10904/10214` [**?**]

In order to isolate potential links to datasets from this list, we perform the following filtering workflow. First, we remove links to domains that are scholarly repositories and which obviously do not host data (or which did not host data prior to 2008). These include domains such as `dx.doi.org`, `arxiv.org`, `xxx.lanl.gov`, and `adsabs.harvard.edu`. Removing links to these domains, which are obviously pointers to articles, narrows down the corpus to 26663.

Second, we remove all links which are found in the reference list of an article. While it is entirely possible that authors cite datasets in the same way as they cite bibliographic references, an exploratory analysis revealed that links in the reference section of a paper were, by and large, pointers to articles, preprints, star catalogs, circulars, manuals, and user guides. Therefore, we remove these "reference links", bringing the corpus down to 20767 links.

Third, based on the observation that links to datasets are generally not found at the root of a website hierarchy, we removed links that contain less than 2 forward slashes (other than the two slashes found in the leading "http://"). For example, the link to `http://www.sdss.org` was dropped from the corpus (0 slashes), while the link to `http://www.cfa.harvard.edu/COMPLETE/data_html_pages/data.html` was retained (3 slashes). This final filtering procedure reduces the corpus to 13447 links, which we consider potential links to datasets. [**?**] Some descriptive statistics about this corpus of links is presented in Table

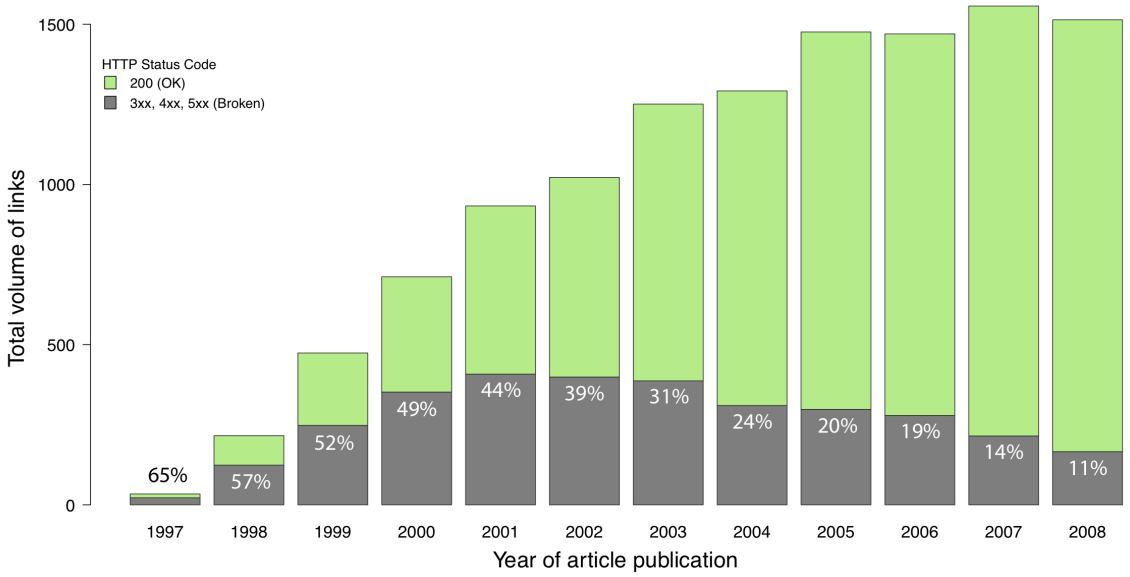

Figure 1. Figure 1. Volume of potential data links in astronomy publications. Total volume of external links in all articles published between 1997 and 2008 in the four main astronomy journals, color coded by HTTP status code. Green bars represent accessible links (200), grey bars represent broken links.

1.

5.1 Acknowledgments

We thank Michael Blake and Tomoko Kurahashi who helped with interviews, transcription, and coding, and with data curation, respectively. We also thank Alberto Accomazzi, Jay Luker, and the Astrophysics Data System team at the Harvard-Smithsonian Center for Astrophysics for providing access to the bibliographic data used for the exploratory data citation analysis.

5.2 Figures

5.3 Tables

Table 1. Table 1. Some descriptive statistics about top domains linked in astronomy publications. This table lists total number of links and broken links (HTTP status codes 3xx, 4xx, and 5xx) to top domains (domains with over 100 links) found within articles published in the four main astronomy journals between 1997 and 2008. The table also shows, for each domain, the portion of links to common filename extensions, as well as links that contain the tilde character.

Domain	links (broken)	.html	.txt	.dat	.gz	.tar	.fits	tilde
cxc.harvard.edu	802 (110)	336 (70)	0	0	4 (2)	5 (4)	1	0
heasarc.gsfc.nasa.gov	640 (33)	423 (27)	1	0	0	0	0	0
www.stsci.edu	498 (61)	205 (29)	3	0	0	0	0	15 (10)
asc.harvard.edu	471 (152)	212 (99)	0	0	0	0	0	1 (1)
ssc.spitzer.caltech.edu	427 (194)	125 (76)	3 (3)	0	0	0	0	0
cfa-www.harvard.edu	352 (68)	277 (52)	1	0	0	0	0	54 (17)
archive.stsci.edu	308 (58)	57 (9)	2	1 (0)	0	0	0	0
www.ipac.caltech.edu	285 (14)	209 (12)	0	0	0	0	0	0
www.atnf.csiro.au	211 (21)	12 (6)	0	0	0	0	0	7 (5)
space.mit.edu	193 (10)	58 (5)	1	0	0	0	0	2 (1)
www.astro.psu.edu	186 (4)	103 (1)	1	10	1	1	0	2
www.eso.org	186 (58)	54 (22)	1 (1)	0	0	0	0	4 (1)
irsa.ipac.caltech.edu	163 (5)	38	0	0	1	0	0	0
www.sdss.org	156 (2)	106 (1)	0	0	0	0	0	0
hea-www.harvard.edu	125 (37)	42 (17)	1	0	0	1	0	26 (16)
physics.nist.gov	125 (3)	63 (2)	0	0	0	0	0	0
www.noao.edu	120 (3)	50 (2)	0	0	0	0	0	0
xmm.vilspa.esa.es	118 (35)	23 (19)	0	0	8 (1)	0	0	1 (1)
www.astro.princeton.edu	115 (31)	43 (14)	0	0	0	0	0	53 (12)
ad.usno.navy.mil	110 (27)	98 (22)	3 (3)	0	0	0	0	1 (1)

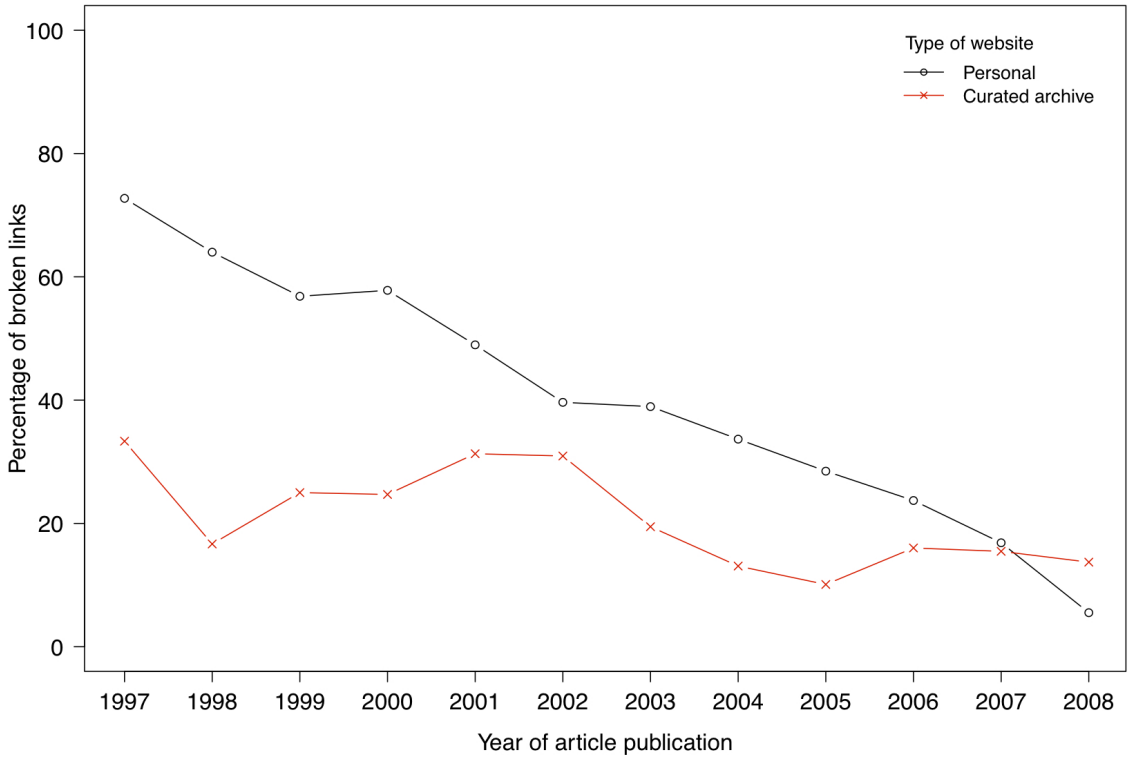

Figure 2. Figure 2. Percentage of broken links in astronomy publications according to type of website. Percentages of broken external links in all articles published between 1997 and 2008 in the four main astronomy journals. Black circles represent links to personal websites (link values contain the tilde symbol, ˜), while red crosses represent links to curated archives such as governmental and institutional repositories.