

Ten Years of SkyServer II: How Astronomers and the Public Have Embraced e-Science

M. Jordan Raddick, Ani R. Thakar, and Alexander S. Szalay | Johns Hopkins University
Rafael D.C. Santos | National Institute of Space Research (INPE), Brazil

A comprehensive analysis of 10 years of Web and SQL traffic on SkyServer—the online portal to the multiterabyte Sloan Digital Sky Survey archive—shows the impressive reach of the SDSS to the research community and the public, and provides insight into how methods of e-science are being taken up by the scientific community.

This article and its companion (published in this same special issue, hereafter referred to as part I) examine Web traffic to a successful e-science project: the online archive of the Sloan Digital Sky Survey (SDSS). The SDSS is an astronomy project launched in 1998 to make a map of the universe. The SDSS makes its entire data archive available online to researchers and the public through a variety of interfaces. The SDSS and its science archive are described in a previous issue of *Computing in Science and Engineering*.¹

This article analyzes 10 years of the SDSS's weblog and SQL log data to understand how scientists and the public are using the e-science resources provided by the SDSS. This work builds upon the analysis previously done on the SDSS log data for a five-year period ending in 2006,² using the normalized database of Web hits and SQL queries described in part I. The general motivation for this type of study, and the uniqueness and significance of the dataset that we have analyzed, are also described in part I.

This current article has the following specific goals:

- repeat all the analyses conducted in the five-year study on the full 10-year dataset;
- characterize the time evolution of Web traffic;
- assess the relative interest in data products and services over time—in particular, how much interest remains in older datasets once new ones have been released; and
- characterize the user population based on patterns in IP addresses and browser use—in particular, how much Web traffic arises from professional

scientists, teachers in the formal education system, citizen scientists from websites like Galaxy Zoo,³ and so on.

We hope that this 10-year analysis will serve as a springboard for further research into how scientists are adapting to the big data paradigm. As described in part I, the SkyServer weblog data are available online, and we encourage other groups to use them to investigate a variety of exciting questions.

Results: Web Traffic

This article focuses on one aspect of SDSS data access: the SkyServer website (see <http://cas.sdss.org>), which offers browser-based access to SDSS catalog data.

For the period we studied—June 2001 through January 2011—the SkyServer weblogs recorded 630 million hits and 357 million page views (see part I for how we define page views as a measure of user volition). The hits originated from 3.4 million unique IPs in 38,508 domains (as resolved by Whois), corresponding to more than 1,000 unique parent organizations that manage these domains, such as universities or ISPs.

The overall statistics for operations (HTTP methods) are 96 percent GET, 3 percent HEAD, and 1 percent POST. In addition to the HTTP methods defined by W3C (see www.w3.org/Protocols/rfc2616/rfc2616-sec9.html), there were a handful of others that were most likely entered as security attacks.⁴

Evolution of Traffic with Time

When the SDSS issued its Early Data Release in 2001, e-science methods were still new; by 2011,

the scientific community had widely adopted these methods. The history of adoption of e-science methodology should be reflected in the usage history of SDSS websites; therefore, understanding how usage of SkyServer has increased over time can give some insight into the increasing usage of e-science methods over a 10-year period.

To examine the growth of SkyServer traffic, we look at the number of page views as a function of time. The 5-year study had analyzed the growth in monthly traffic to SDSS websites;² here, we analyze the hits by day; however, our results are directly comparable to those in the previous study. The blue dots in Figure 1 show page views to SkyServer per day.

As is common in Web analytics, traffic per day shows wide variability over short timescales. We smoothed the daily page views using a moving average smoothing, trying different smoothing periods before settling on a 30-day window, which has the added advantage of making our results directly comparable with the monthly results in the previous 5-year study. In Figure 1, red dots show smoothed page views per day.

One obvious feature in Figure 1 is the sudden peak in usage in mid-2007. This peak is due to the launch of the Galaxy Zoo citizen science project,³ which drove hundreds of thousands of people to the website (now archived at <http://zoo1.galaxyzoo.org>) to view and classify SDSS galaxy images.

In terms of longer-term trends in use, the five-year study reported that traffic had doubled each year from 2001 to 2006. Figure 1 shows that this doubling behavior didn't continue through the entire 10-year period, but because of short-term variability it's difficult to tell exactly when traffic switched to a steady state. To try to pinpoint this time, we fit the observed daily page views with a simple (probably too simple) model in which page views increase exponentially until a particular day, at which point the number per day became constant until the end of the study period.

We allowed that transition date to vary and found the exponential growth rate and steady-state value that best fit the data, using a Kolmogorov–Smirnov (K-S) test.⁵ In the best-fitting model ($R^2 = 0.699$), page views to SkyServer grow exponentially, doubling about every year, until August 2007, when they level off at about 200,000 per day.

There are also unsurprising daily and monthly regularities in SkyServer traffic. SkyServer receives significantly fewer page views (about 20 percent less) on weekends than during the week, and the highest traffic occurs on Wednesdays. Page views

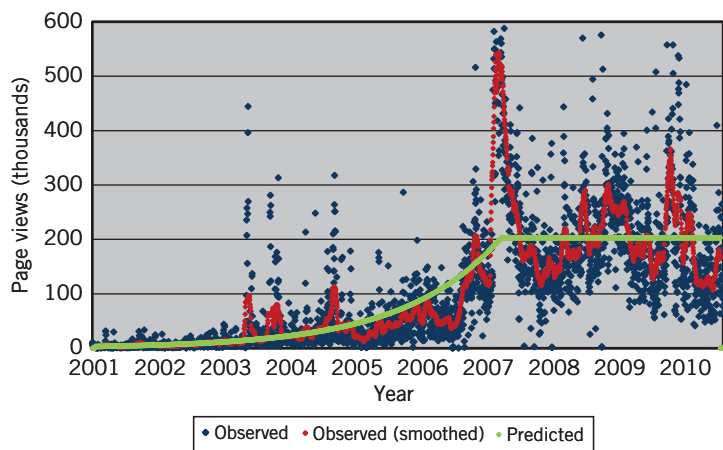


Figure 1. Page views to SkyServer per day (blue), smoothed using a 30-day moving average smoothing (red), and compared to the best fitting model of exponential growth followed by steady state (green).

generally increase during the academic year, up to a peak in May, then fall markedly during the summer months.

Characterizing the User Population

An obviously important question in Web analytics is, who is using my website? A page view in the SkyServer weblog includes three pieces of information that can help categorize what type of user viewed that page: the user agent (the browser, program, and so on, that they used to access the site), the client IP address (what Internet address their request came from), and the Web commands they executed (generally equivalent to what pages they requested).

The following sections analyze hits and page views to SkyServer in terms of client Web agents and in terms of client IPs and domains. Next, we discuss which data and pages are most often requested from SkyServer.

User Agents

Each hit record in a weblog contains an agent string that identifies what type of browser or program is generating the request (for example, Microsoft Internet Explorer, Mozilla Firefox, Google search engine robot, wget data downloader, and so on). However, some agents don't report agent strings, and others falsely report to be popular browsers such as Internet Explorer in hopes of bypassing firewalls or obtaining some other desired result.

Most hits to SkyServer (approximately 60 percent) come from Web browsers, indicating a site that receives most of its use from human users. However, a significant amount of traffic (approximately 30 percent of

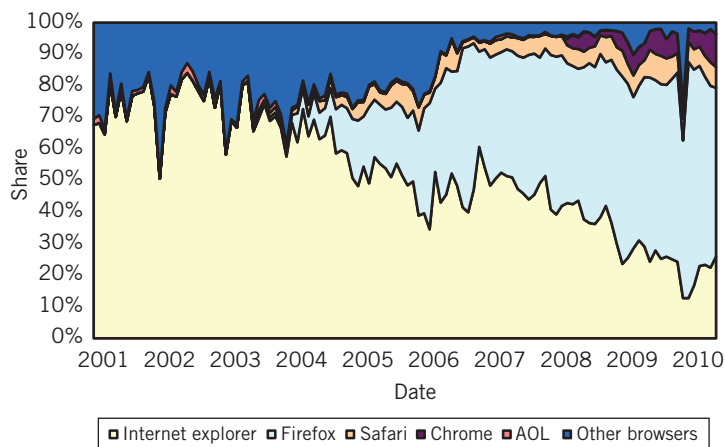


Figure 2. Percent share of page views by Web browser by month from July 2001 to December 2011 for each of five popular browsers, plus others.

hits) does come from programs, indicating that users also write scripts or programs in various languages to download data on their behalf.

What is the nature of programmatic access to SkyServer? Agent strings, when reported, show which program accessed the site on the user's behalf. Thus, interpreting Web agents can provide some insight into the changing ways in which researchers automate their access to scientific data.

Of hits that came from agents classified as programs, Python was the most popular language, with 61 percent of total hits. Following were wget (24 percent), Java (6 percent), Perl (5 percent), and World Wide Telescope (4 percent), with others accounting for less than 1 percent. Python became the most-used programmatic agent starting in mid-2003, and now accounts for more than double the hits from the next most popular programmatic access mode, wget.

The most popular way to access SkyServer was through a Web browser. This is as we would expect—SkyServer was designed to provide simple, browser-based access to the most common SDSS science parameters. Figure 2 shows the percentage share of browsers in the SkyServer weblogs for five popular browsers and others. Page views began appearing from Firefox in early 2005; by early 2009, Firefox had become the most popular browser used to view SkyServer.

Comparing the data in Figure 2 to statistics on browser share from 2008 to 2011 from Stat Counter (see <http://gs.statcounter.com/#browser-ww-monthly-200807-201310>) shows that SkyServer users are more likely to use Firefox and Safari, less likely to use Internet Explorer, and about equally likely to use Google Chrome as Web users overall.

Inferring User Identity with Client IPs and Domains

Although we can't know the identity and intentions of a user from Web analytics, we can identify the IP address from which the user's request originates—and this information lets us make some educated guesses about those larger questions.

IP addresses are assigned hierarchically in ranges; the entity that owns a particular IP range is known as the domain. Many organizations own multiple domain ranges, sometimes under names that are only slightly different (for example, "AT&T Internet Services" versus "AT&T Services, Inc."), meaning that some human judgment is required to identify what parent organization a particular domain name belongs to. The way in which we go from IP addresses to resolved domains to domain parent organizations is detailed in part I.

This means that, in most cases, the IP address recorded in the SkyServer weblogs can be associated with a specific organization through its domain name. Knowledge of these institutions won't always prove helpful in inferring user intentions—for example, traffic from a home ISP such as AT&T or Brasil Telecom might indicate any of a large number of user scenarios. But sometimes, such as with traffic from large research institutions or K–12 schools, inferences of user goals can indeed be made.

Client IP addresses. During our study period (2001–2011), SkyServer received at least one page view from 3,081,621 valid IP addresses. The number of hits per IP address was highly uneven. The IP address with the most traffic was a client at the University of California, Berkeley, with more than 10 million page views during the study period. At the other end of the distribution were the more than 1 million IP addresses that had only a single page view (35 percent of all IP addresses had only one page view).

Our five-year study had found a power law relationship in the number of hits per IP address, but using methods recommended in the literature,⁵ we found that a power law wasn't a good fit to the data. But regardless of the specific form of the relationship, it's clear that SkyServer traffic follows a pattern typical on the Web: dominated by many users with only few accesses each, but with a significant long tail in which some users have deep and repeated interactions.

Client domains and organization. IP addresses are highly useful in estimating how a site's traffic varies among different users, but they aren't the most useful

unit of analysis for making more specific inferences about who those users are, for two reasons. First, IP addresses belong to computers, not people—if a user moves to a different computer, his or her IP address will change. Second, many routers and ISPs assign IP addresses using the Dynamic Host Configuration Protocol, which assigns a different IP address to the same computer at different times.

A way to circumvent some of these problems and draw additional conclusions about website users is to look at Web traffic at the level of domains and their parent organizations, rather than individual IP addresses.

Through an iterative process of scripted Whois calls followed by closer examination of returned domains, we were able to identify domain names for 93 percent of the IP addresses in our weblogs, and domain parent organizations for 71 percent. This process is described more fully in part I.

Hits to SkyServer resolved to 37,597 separate domain ranges, associated with 1,047 parent organizations. As with individual IP addresses, the number of hits per domain was highly uneven.

One important piece of information contained in the Whois results is the country of that Web domain (see Table 1). SkyServer received at least one page view from 160 different countries. The largest number of page views (43 percent) came from domains in the US—although we suspect that because many international Web users use ISPs based in the US, this number likely underestimates the true international reach of SkyServer. At any rate, the distribution of page views by country is highly uneven, with just 18 countries accounting for 95 percent of SkyServer page views.

Table 2 shows the 10 domain organizations from which the most SkyServer page views originated. What are these organizations? We divided the organizations into categories for which we might expect meaningful differences in usage patterns. Results are shown in Table 3. The largest number of hits came from commercial ISPs. It's difficult to draw conclusions from this fact, because many different types of users might use commercial ISPs (for example, researchers working from home, high school students doing homework, and so on), but we're confident that this indicates usage of SDSS data resources by the general public. Clearer categorization is likely possible through a combination of domain organization, Web browser, and pages accessed.

Of non-ISP page views, most come from research institutes (for example, Space Telescope

Table 1. SkyServer page views by country, as indicated by the domain name "country" field returned by a Whois lookup. These 18 countries account for 95 percent of SkyServer page views.

Country	Page Views (%)
US	42.7
European Union	16.9
Australia	7.7
United Kingdom	7.1
Germany	3.8
Canada	3.3
South Korea	2.3
Japan	2.2
Netherlands	1.8
China	1.4
Spain	1.0
Italy	1.0
Brazil	1.0
Mexico	0.9
Chile	0.7
France	0.6
Hungary	0.5
Poland	0.5

Table 2. The 10 organizations with the largest numbers of SkyServer page views.*

Organization	Page views
Harvard-Smithsonian Center for Astrophysics	12,975,691
University of California	10,687,306
Fermi National Accelerator Laboratory	10,395,671
AT&T	10,133,291
Comcast	9,540,580
RoadRunner	4,903,561
Space Telescope Science Institute	4,754,897
BT Internet	4,736,879
Japan Aerospace Exploration Agency	3,804,890
National Research Council Canada	3,680,220

*This table shows the organization and the number of page views from that organization.

Science Institute) or universities, but a few percent of page views come from colleges, community colleges, or K–12 institutions (for example, schools, districts, and statewide educational networks).

One limitation of this analysis is that we don't account for changes in IP domains, for example, when

Table 3. SkyServer page views by type of organization.*

Category	Organizations	Page views (%)
Research institute	92	25.3
University	295	35.2
College	14	0.3
Communications/Technical college	9	0.1
K-12 institution	57	0.8
ISP	494	36.1
Business (not ISP)	55	1.2
Government organization	31	1.0

* This table shows the number of organizations (out of 1,047 total), and the percentage of page views per organization type.

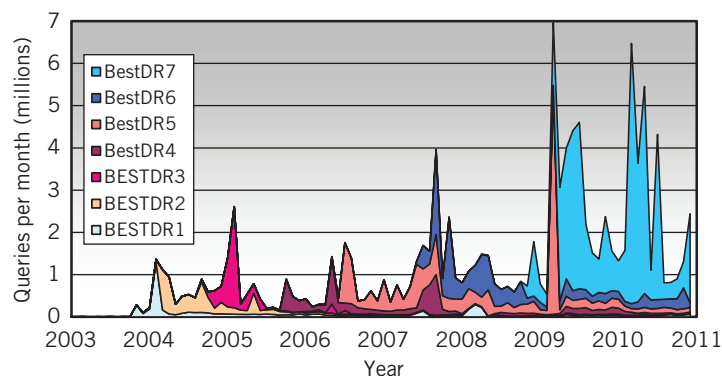


Figure 3. SQL queries per month to each data release, showing how interest in each release changes as newer data becomes available.

an ISP is bought by a larger company; we ran our Whois scripts in 2012 and used the results to draw conclusions about Web traffic from 2001 to 2011. While this means we have undoubtedly misclassified some domains, it's unlikely that many domains changed country or domain type (for example, from an academic institution to a commercial one). Therefore, we're confident that this effect has little impact on our overall conclusions.

Traffic by Data Product

SkyServer provides access to all the data releases of the SDSS, not just the most recent. This is by design: we wanted to make it as easy as possible for scientists to reproduce results of earlier work using prior data releases. This raises the question of how extensively are these prior data releases used.

We experimented with several ways of measuring use of prior datasets directly from weblog data, but we were unable to find a good metric. Instead, we turned to the SQL query logs to see which databases users are querying. The SQL logs retain the

full text of each query submitted to SkyServer since 2002. Because they record the database to which each query was submitted, they provide a much more reliable and accurate measure of which users are looking for which datasets.

Although SkyServer hosts many specialized datasets, nearly 98 percent of traffic goes to the best SDSS data release, BestDR n , where n is the release number. During our study period, the SDSS completed eight data releases, from the early data release to the seventh data release (DR7). Figure 3 shows the change in the total number of queries to various data releases over the period of this study. Note the level of interest in old data releases: for example, BestDR4 was still receiving an appreciable number of queries in 2010, five years after its release.

By the end of the study period in 2011, the latest release (DR7) dominated the SQL query traffic, particularly because this was the last public release from the second phase of the SDSS project, and was largely seen by the community as a completion of the original SDSS survey mission. Older releases see their queries drop off sharply when new ones are released, but importantly, they keep that reduced level of traffic for years as the queries to them taper off slowly over time.

Requests by Page

Here we detail the patterns of usage of specific pages and subsections of the SkyServer website. Our logs contain the Web command that was executed by the server; in most cases, this is the page that the user requested. We split the command around the “?” symbol into two parts: the stem, which gives the page requested; and the parameters, which give the options that were passed into that page. We conducted the analysis in this section using command stems.

In terms of page types, more than 75 percent of hits and 99 percent of page views go to SkyServer's main content page types: .asp and.aspx. In terms of the site's navigation structure, more than 92 percent of page views go to the data access tools section of the site. The next most popular section, with 3 percent of page views, is the educational projects. This section discusses the usage of those two subsections.

Traffic by data access tool. SkyServer offers a wide variety of data access tools, available by clicking on data access links on the homepage. Growth in traffic to these tools pages has followed the same pattern as growth to the overall site—page views in the

tools directory grew from 900,000 per year in 2002 to more than 40 million per year in both 2009 and 2010. Figure 4 shows the distribution of page views by data access tool. The tools, in order of decreasing number of page views in our dataset, are as follows:

- *Search*. Users can search for sky objects using a variety of criteria, from simple position-based searches to user-created, raw SQL queries.
- *Visual tools*. Users can view data superimposed on sky images in various ways, including interactive images.
- *Explore*. Users can view all information available in the database for a single sky object.
- *Scrolling sky*. Users can view sky images by scrolling, simulating the way in which the SDSS telescope’s drift scanning sees the apparent motion of the sky each night.
- *Quick look*. Users see a simplified view of the explore interface with just the most frequently requested columns of data, ideal for students and teachers learning astronomy.

The largest category of tools (in terms of page views) are the SkyServer search tools, which let users search the entire SDSS database for objects that meet criteria they specify. Figure 5 shows the distribution of page views within the category of search tools. A number of different search methods are available, ranging from a simple search by position for novice users to a raw SQL interface for experts.

Surprisingly, the raw SQL interface is the most popular method of searching SkyServer, suggesting extensive use from a population of users who are comfortable with both computer science and astronomy concepts.

Traffic by educational project. One of the unique features of SkyServer is a series of educational projects that use SDSS data to teach concepts in astronomy, science, mathematics, and other fields. The projects are available under the “education” links on the home page. The English-language project pages received a total of 6,410,064 page views during the study period, making up 3.3 percent of the total page views to the English-language SkyServer site. However, all the projects require learners and instructors to go out from the projects pages to the tools pages to retrieve data; therefore, the number of page views to the projects directory should be considered a lower estimate only.

The following are SkyServer projects main categories:

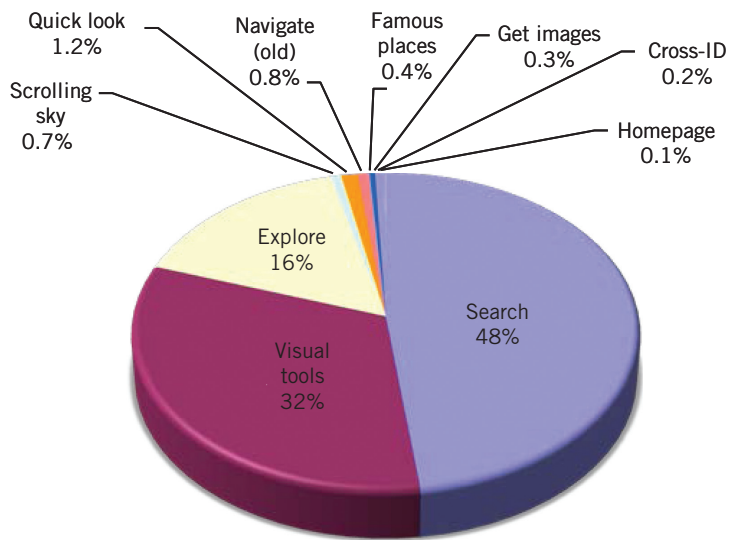


Figure 4. The usage of various data access tools is profiled in the pie chart. The search tools account for the majority of the page views, with the visual tools for image browsing being the second most popular means of accessing the data.

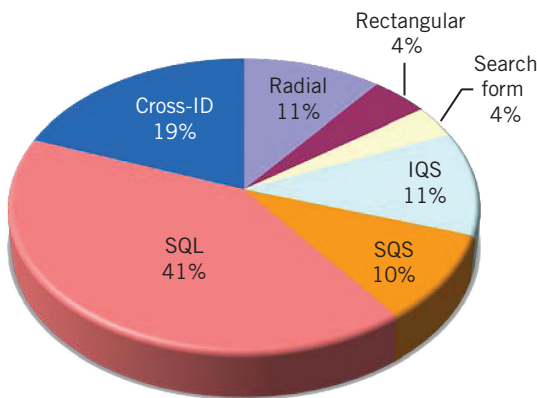


Figure 5. Distribution of page views by search tool. The raw SQL search is the most popular search tool. (IQL = Internet Query Language; SQS = Simple Queue Service.)

- *basic*—appropriate for high school science and introductory college (for example, Astronomy 101) courses;
- *advanced*—appropriate for advanced high school science, technology, engineering, and mathematics, and honors-level college courses; and
- *for kids*—appropriate for upper-level elementary school and middle school students.

We deliberately chose to categorize projects in this way rather than by assigning specific grade levels, so that instructors at a variety of different levels could use them flexibly.

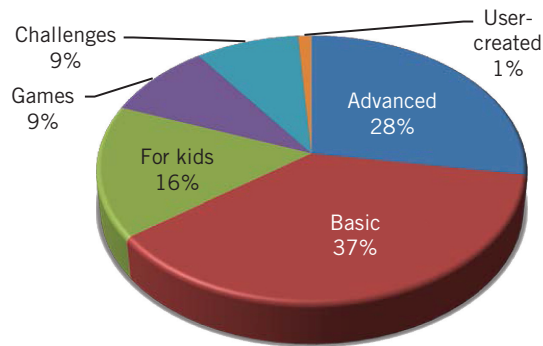


Figure 6. Distribution of page views among educational project pages. Each level of project receives a significant amount of traffic.

These projects include complete teacher guides that give sample solutions, advice on leading a class through the project, and lists of what national science standards are met by each project. The teacher guide pages received 459,605 page views during the study period.

In some cases, the different levels contain different versions of the same projects, with content and activities tailored to each user level. In other cases, different topics are covered for the different levels. Thus, there are different numbers of projects for each level: nine for advanced, but only five for basic. Therefore, a fair comparison of page views among the various project levels would normalize the total page views by the number of projects in each level. Figure 6 shows the normalized percentage of page views at each project level during the study period. It shows that each level of project receives a significant amount of traffic.

Within each level, which projects are most often used? Again, different projects have different numbers of pages, so the best comparison would require normalizing by the number of pages per project. Figure 7 shows the percentage of page views for each individual project in both the basic and advanced project categories. Stars are the most popular type of object in either category, with galaxies and cosmology a close second. Figure 7 shows that educators are using SkyServer projects to teach a wide variety of topics in astronomy and other sciences.

This article has conducted a basic first-order analysis of the SDSS database to find patterns in how scientists and the public use e-science resources. Our analysis focused on page views as a measure of con-

sciously-chosen user activity—defined as error-free hits from browsers or programs to content pages.

We began our study by repeating prior analyses done on 5 years of SkyServer weblog data. The biggest difference we found when looking at the full 10-year data was that the exponential growth identified previously ended around the beginning of 2008; since then, page views to SkyServer have varied around a steady-state average of about 200,000 per day. We also found that the distribution of hits by IP address and domain didn't show a power law behavior in our larger 10-year dataset, but is still highly uneven, with a few IPs accounting for most traffic and a long tail of IPs with a few page views each. This pattern is typical of websites and other online resources. Predictably, our power users come from the scientific research community (more than 60 percent, but this is most likely a gross underestimate because we can't characterize ISPs), but a significant fraction of our data access comes from nonscientific users worldwide who are large in number but who don't make frequent or intensive data requests.

We looked at page views in terms of time, user agent (for example, browsers), IP address, and domain, and command executed (for example, page visited). About 70 percent of page views come from Web browsers and about 30 percent come from programs that users write to download data on their behalf. The latter are almost entirely professional astronomers for whom our data is integral to their research. Page views come primarily from universities, research institutes, and commercial ISPs, although there is significant usage in K–12 schools and small two- and four-year colleges.

Logs of SQL queries submitted to SkyServer show a continuing interest in data products even many years after those products have been superseded by more recent products. This observation vindicates our choice to always retain all data, and we encourage other data providers to do the same. The most popular data access method on SkyServer is the raw SQL interface, indicating that the new methods of data-intensive science have gained a significant foothold in astronomy research. An important conclusion from this finding is that Web-form queries and other graphical query tools are only useful up to a point, and eventually users will gravitate towards the query tool that provides the most versatility and direct access to the data (that is, the raw SQL interface), even if it involves a significant learning curve (provided there are enough help resources onsite).

SkyServer's educational projects have also received significant use at both the Basic and Advanced

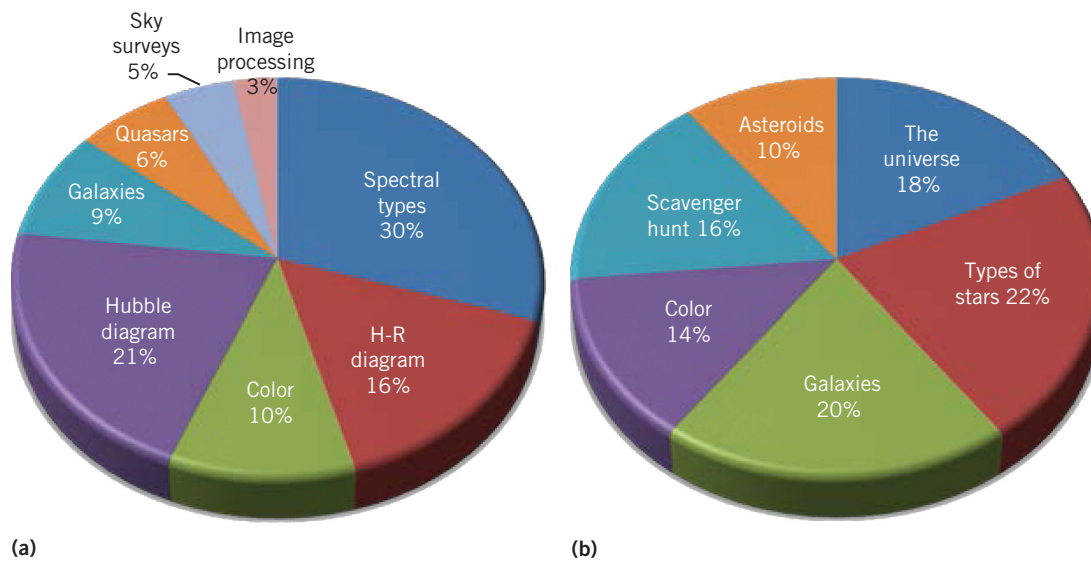


Figure 7. The distribution of page views in each of the (a) advanced and (b) basic educational projects. Types of stars invoke the maximum interest in both categories.

levels, with content throughout the astronomy curriculum proving popular. This observation, coupled with the significant traffic from K–12 schools as well as the traffic driven by Galaxy Zoo, indicates that SkyServer has done an excellent job of serving the needs of the general public.

It would be useful to compare these results for SkyServer to similar results from other websites, both those that host scientific data and in general; however, few sites publish this information with enough detail for comparison. To our knowledge, this is the only study of its kind to date in the literature, where 10 years of online access data for a multiterabyte public data set is analyzed to study who’s using it, how it’s being used, and especially how the new tools of e-science are being received by the community.

And yet, the analysis presented here is only the beginning. A much more informative analysis can be done by combining two or more of these variables presented here. For example, how are patterns of data access similar or different among users originating from IPs associated with research universities, small colleges, and K–12 schools? Can we use browser version statistics to identify users at various stages of technology adoption according to the diffusion model,⁶ and what differences in behavior exist between, for example, early adopters and the majority?

An additional set of questions can be answered by considering the logs of SQL queries submitted to SkyServer. For example, the ordered history of

SQL queries submitted from the same IP address is likely to reveal a pattern of progressively more complex queries submitted in a trial-and-error process, which will provide insight into how users learn SQL in a naturalistic setting, and potentially help improve systems for teaching SQL and other information technology skills.⁷ Additional insights into these questions can be gained by reconstructing user sessions and studying how users navigate back and forth between accessing webpages and submitting SQL queries.

We’re currently investigating these questions and others, and we welcome collaboration with other groups in this exciting venture. We hope that SkyServer weblogs will help the e-science community grow and develop through better understanding its users. ■

Acknowledgments

This material is based upon work supported by the Alfred P. Sloan Foundation.

References

1. A.R. Thakar, “The Sloan Digital Sky Survey—Drinking From the Fire Hose,” *Computing in Science & Eng.*, special issue on SDSS science archive, vol. 10, no. 1, 2008, pp. 9–12.
2. V. Singh et al., *SkyServer Traffic Report—The First Five Years*, tech. report MSR-TR-2006-190, Microsoft Research, 2006; <http://research.microsoft.com/apps/pubs/default.aspx?id=64520>.

3. C.J. Lintott et al., “Galaxy Zoo: Morphologies Derived from Visual Inspection of Galaxies from the Sloan Digital Sky Survey,” *Monthly Notices of the Royal Astronomical Soc.*, vol. 389, no. 3, 2009, pp. 1179–1189.
4. R.D. Santos et al., “Analysis of Web-Related Threats in Ten Years of Logs from a Scientific Portal,” *Proc. SPIE, Defense, Security, and Sensing*, vol. 8408, 2012, p. 84080H.
5. A. Clauset, C.R. Shalizi, and M. Newman, “Power-Law Distributions in Empirical Data,” *SIAM Review*, vol. 51, no. 4, 2009, pp. 661–703.
6. E.M. Rogers, “A Prospective and Retrospective Look at the Diffusion Model,” *J. Health Communications: Int’l Perspectives*, vol. 9, no. 1, 2004, pp. 13–19.
7. A. Mitrovic, “Fifteen Years of Constraint-Based Tutors: What We Have Achieved and Where We Are Going,” *User Modeling and User-Adapted Interaction*, vol. 22, nos. 1–2, 2012, pp. 39–72.

M. Jordan Raddick is a science education developer at Johns Hopkins University. He’s the science evangelist for the Sloan Digital Sky Survey, focused on making SDSS data available and useful to the scientific community and the world. His research interests include looking at how citizen scientists learn science by doing science. Raddick has an MS in science writing and education from the Johns Hopkins University. Contact him at raddick@jhu.edu.

Ani R. Thakar is a principal research scientist at Johns Hopkins University. His research interests include data-intensive science and interacting galaxies. Thakar has a PhD in astronomy from the Ohio State University. Contact him at thakar@jhu.edu.

Alexander S. Szalay is the Alumni Centennial Professor of Physics and Astronomy at Johns Hopkins University. His research interests include cosmology, the large-scale structure of the universe, data mining, and science with large databases. Szalay has a PhD in astrophysics from Eötvös University, Hungary. Contact him at szalay@jhu.edu.

Rafael D.C. Santos is a senior technologist at the National Institute of Space Research (INPE), Brazil. His research interests include data mining and distributed systems. Santos has a PhD in computer science from the Kyushu Institute of Technology in Japan. Contact him at rafael.santos@inpe.br.



Selected articles and columns from IEEE Computer Society publications are also available for free at <http://ComputingNow.computer.org>.