SLAC-TN-12-017

INSPIRE and SPIRES Log File Analysis

Cole Adams
Science Undergraduate Laboratory Internship Program

Wheaton College

SLAC National Accelerator Laboratory

August 5, 2011

Prepared in partial fulfillment of the requirement of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship under the direction of Travis Brooks in the Scientific Computing division at SLAC National Accelerator Laboratory

Participant Signature: _____

Research Adviser Signature:_____

Work supported by US Department of Energy contract DE-AC02-76SF00515.

ABSTRACT

SPIRES, an aging high-energy physics publication data base, is in the process of being replaced by INSPIRE. In order to ease the transition from SPIRES to INSPIRE it is important to understand user behavior and the drivers for adoption. The goal of this project was to address some questions in regards to the presumed two-thirds of the users still using SPIRES. These questions are answered through analysis of the log files from both websites. A series of scripts were developed to collect and interpret the data contained in the log files. The common search patterns and usage comparisons are made between INSPIRE and SPIRES, and a method for detecting user frustration is presented. The analysis reveals a more even split than originally thought as well as the expected trend of user transition to INSPIRE

INTRODUCTION

SPIRES is the high-energy physics publication database that has been used for collaboration and communication between physicists since the late 1960s. However, forty years is a long time in terms of technology, and SPIRES is in the process of being replaced by a new, similar service called INSPIRE. INSPIRE builds off of the usefulness of SPIRES and hopes to provide the high-energy physics community with better tools for successful research. INSPIRE provides a much faster alternative to SPIRES with a greater depth of features. INSPIRE functions as the quality SPIRES content built on the modern Invenio digital document repository platform developed at CERN

While INSPIRE is an incredibly high-quality tool, it is not yet a finished product. Currently INSPIRE is in a public beta phase and it will become a full production service in the Fall of 2011. Any service can use improvement and the best way to ensure that this improvement is most beneficial is to get feedback from the users themselves. Having a user base that provides constant feedback about the website will give the developers a clear focus of what will help the users most.

At the moment, the user base is split between SPIRES and INSPIRE, and originally it was thought that roughly two-thirds of the users still were using SPIRES, based on a simple

metric of total searches. There were two questions to be answered about these two groups:

1. What are the distinctions between the two groups?
2. Is the two-thirds division the correct one?

In order to answer these questions we analyzed the log files for user behavior to find a better usage metric and to describe the differences between user behavior in INSPIRE and SPIRES. The methods used in this project revealed the presence of a large number of robots that artificially increased the search counts on SPIRES, making the original usage metric a poor one. Using a new metric based on user session results in a division of almost 50-50. The original question about the distinction between these two groups is still important because the SPIRES users will need to switch to INSPIRE soon. There is an effort to try and get the remaining SPIRES to switch over to INSPIRE before SPIRES is shutdown so that these users can make the change without feeling forced, a situation that can create unnecessary work and pain for both users and developers alike.

METHODS

My part in the INSPIRE project is to build statistical tools that can assist developers in deciding how to encourage SPIRES users to make the switch to INSPIRE now, as well as determining functionality that INSPIRE lacks which can then be added. By looking at the usage patterns of both SPIRES and INSPIRE the developers can see what features users most want to retain from SPIRES and what features they use the most in INSPIRE. From that information the developers can guide the SPIRES users to those features in INSPIRE. It is also important to find what barriers prevent the switch to INSPIRE. There are a variety of possible reasons, such as a lack of knowledge about INSPIRE or possible missing features in INSPIRE. We can learn what

the barriers might be through analysis of the usage patterns.

In order to learn about the usage patterns of SPIRES and INSPIRE, there are two ways to learn about the users. One way is to have the users continually submit feedback. This system is already in place and allows for the developers to communicate with their users. The second is to look through the log files and examine the search queries and site hits that the users have generated through using the website. This alternative method had only been implemented at a basic level, and it was the aim of this project to improve the depth and functionality of this technique.

Some of the data that exists in these log files that can be analyzed are session lengths, common search terms, search results, and geographical or institutional location of users. This information is useful to the developers in the transitional period between SPIRES and INSPIRE, and will continue to be important as they try to keep improving user experience on the website. It can directly help identify the barriers mentioned above; common search terms can show the differences between search patterns in SPIRES and INSPIRE, location can tell us if some regions are possibly unaware of the impending switch, and search results can let us know if people are finding what they are looking for, and the session information can tell us if people are either getting frustrated with INSPIRE or if they are pleased with the performance of the site. The data extracted from the logs can also be used to build more complex data such as reconstructing complete user sessions. The user session information can be used to track things such as user frustration, when a user is struggling to use a part of the website effectively. The administrative staff of INSPIRE could then see where this occurs and send a message to the user to assist them. Analysis of the logfiles may reveal additional uses of the logfile data.

There is a large amount of data associated with the log files resulting from over 15,000

searches performed by users every day. Since the log files contain every HTTP request to the website, there is a lot more data than just searches in the logs, and over time the data that needs to be processed can become fairly large. It would be extremely inefficient to process the data from scratch every time a different analysis needs to be done, so the data is cached at various stages to improve performance. This caching allows months worth of data to be processed in minutes.

The log file analysis was performed using Python, a free scripting language. The analysis contained in this paper was all done with a set of log files from May 1, 2011 through July 31, 2011; however, the analysis tools were designed with the intention of long term usage that will be able to repeat the analysis later. The most useful methods can potentially be added to some of the existing basic log file statistics that are already in place. This will make the analysis techniques more easily maintainable.

While analyzing some of the data early on it became apparent that there were a number of IP addresses from China that had an enormous number of searches. Upon further investigation, the searches they were performing appeared very scripted. These were considered to be robots and were not considered in further analysis.

RESULTS

Search Keyword Occurrence

The search keyword occurrence was extracted by examining each search query for any keywords or their aliases. For instance, 'au', and 'author' both map to the search keyword 'author'. The top ten search keywords for both INSPIRE and SPIRES are shown below in Table 1.

INSPIRE (June 2011)		SPIRES (June 2011)	
Search Keyword	Occurrences	Search Keyword	Occurrences
'author'	259243	'author'	278062
'field'	58225	'date'	251327
'title'	39090	'date-added'	237226
'date'	25316	'reportnumber'	89219
'reportnumber'	13824	'irn'	67713
'journal'	13512	'exactauthor'	65081
'exactauthor'	9801	'reference'	64565
'keyword'	4367	'title'	38420
'collaboration'	4300	'journal	25183
'reference'	4165	'keyword'	19672

Table 1: SPIRES Search Keyword Occurrences in INSPIRE and SPIRES

User Sessions

The user session data contains the searches that a user performed during a visit to either INSPIRE or SPIRES. The session data is constructed by looking at searches performed by a given IP (which is assumed to identify a unique user) and uses the time in between successive searches to decide whether or not the search reflects the start of a new session or a continuation of the current session. Both the time of the search and the search query are recorded for each search in a session. Figure 1 shows a histogram of the session searches for both INSPIRE and SPIRES. The data from both follows the same shape when the data is scaled to match the second points of both logs, with the exception of the first data point (which represents the single search session counts). For this reason, the multi-search session counts were considered distinctly from total session counts. A multi-search session represents a countable use of the website, so the multi-search session counts is taken as a superior metric to the total search counts.

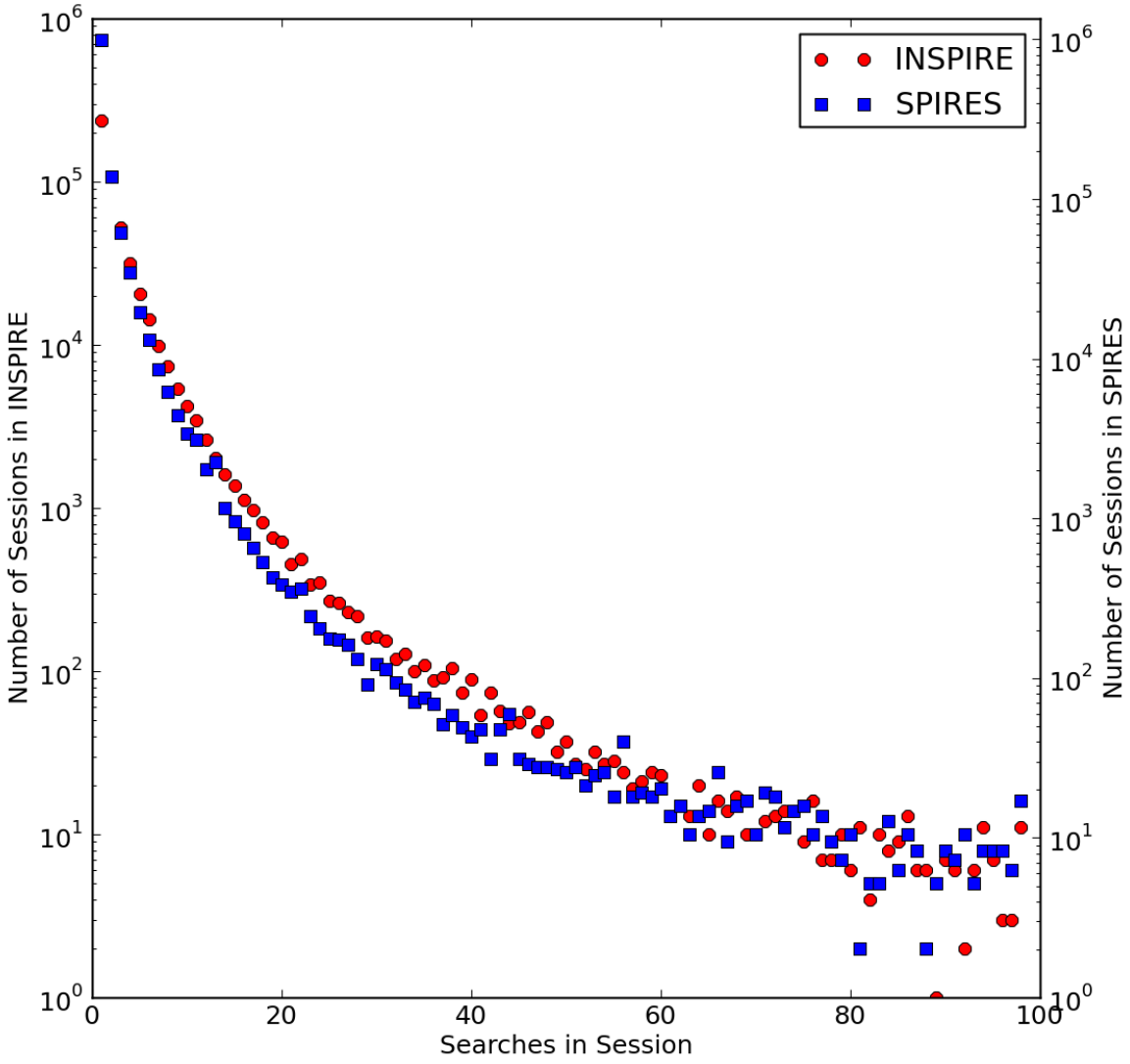

Figure 1. Histograms of the number of searches in a session.

Usage Statistics

Based on the session data shown in Figure 1, the two best measures of usage for INSPIRE and SPIRES are the number of unique users and the number of

multiple search sessions. Figure 2 shows a plot of usage metrics over this time span.

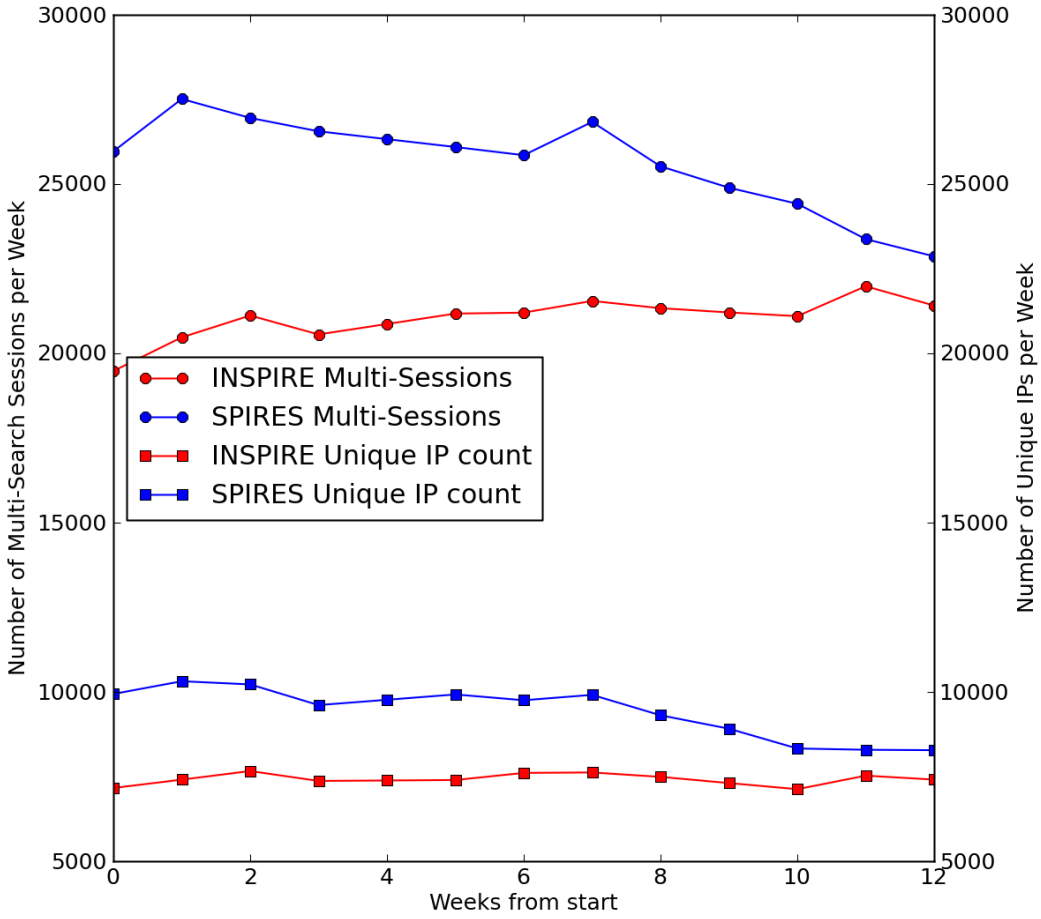

Figure 2. Plot of usage metrics (Multiple Session and Unique IP counts) over time

The usage can be broken down by country to show where the use of INSPIRE and SPIRES is most prominent as well as to give a good idea about where INSPIRE has been adopted in place of SPIRES. Figure 3 shows a bar plot of of the multiple search sessions in INSPIRE and SPIRES as well as the users who used both INSPIRE and SPIRES, broken down by their session count in each website. The segment labeled “Others” refers to any country that

made up less than 5% of the whole in all of the categories.

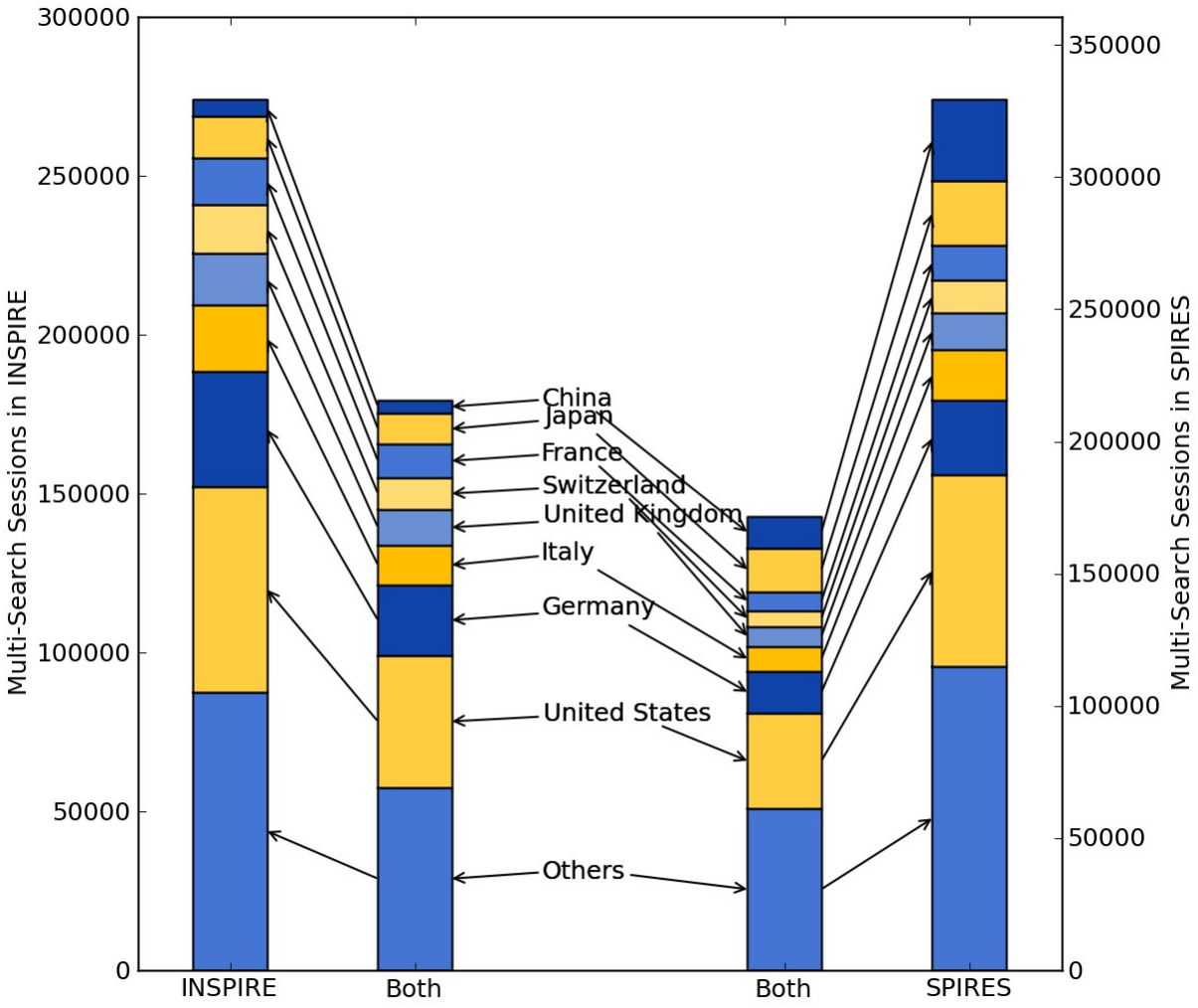

Figure 3. Bar plot of usage broken down into countries.

User Frustration Analysis

During a user's session it is possible that the user can experience a “frustration event”, where the user cannot seem to find the correct way to execute a given search query. Both SPIRES and INSPIRE use fielded metadata searches, which means that they have syntaxes that allow the users to specify the type of information they are searching on. This is useful in physics for searches such as “Higgs”, which is both an author and a frequent word in a title. When the

search syntax doesn't work as expected it can lead to user frustration. This shows up in the logs as several consecutive searches by a user that are similar in structure and content. These frustrated sessions are recorded and divided into two categories based on the severity of the frustration events present in the session. Users in the extremely frustrated category have twice as many repeated similar searches as those in the frustrated category. Table 2 shows the frustration event counts for INSPIRE and SPIRES as well as the total number of sessions that are long enough to potentially contain a frustration event. The frustration events are usually centered on a search keyword, so the breakdown of the most common search keywords that appear in frustration events is shown in Table 3.

	INSPIRE		SPIRES	
	Frustrated Sessions	Total Sessions	Frustrated Sessions	Total Sessions
Frustrated	289	14613	197	13830
Extremely Frustrated	255	5967	220	5656

Table 2. “Frustration Event” counts for INSPIRE and SPIRES

INSPIRE		SPIRES	
'author'	37%	'reportnumber'	29%
'journal'	22%	'author'	26%
'field'	11%	'exactauthor'	15%
'title'	7%	'title'	9%
'reportnumber'	5%	'journal'	9%

Table 3. Common search keywords in “frustration events”

DISCUSSION AND CONCLUSION

The search keyword data shown in Table 1 can give a good estimate as to what INSPIRE is being used for. The high number of author searches demonstrates that the preferred search method is to search by author. It also gives an idea of what search keywords are most used in

SPIRES which can help to direct focus towards ensuring that the search engine in INSPIRE handles them properly. In this case, focus should be directed towards ensuring that the author search works properly.

The session data is highly interesting as a usage metric, particularly when the large variation in the SPIRES single search session is discounted(Figure 1). The two most general explanations are external links and automated systems. In the case of external links, a user is simply clicking a link that returns a SPIRES search page, and no actual search is performed user. In the logs, this would show up as a session with one search, and is not particularly interesting to look at. In the case of automated systems, there is perhaps some script that runs a search once every hour to refresh a list of articles that fits some search query. Both of these would artificially increase the number of single search sessions and would be sufficient reason to discount the point. The reason the same variation is not present in INSPIRE is perhaps due to how new it is relative to SPIRES. In future work it might be possible to identify robots and remove their searches entirely and correct the anomalous single search session count.

The usage data gives some useful results in terms of comparisons between INSPIRE and SPIRES. The usage time line in Figure 2 indicates that SPIRES usage is decreasing, while INSPIRE usage is increasing only slightly. There could be other trends that would occur in a system with constant usage that might mask an increasing trend in INSPIRE which would correlate to the decreasing trend of SPIRES. A seasonal effect that takes place from May 1 to July 31, such as summer vacations, could explain this.

The bar chart in Figure 3 helps break down the usage by the origin of the user. The plot shows that the United States makes up the plurality of both INSPIRE and SPIRES usage, but also that both services are clearly used internationally. The relative sizes of countries in

INSPIRE and SPIRES can identify where people have begun transitioning to INSPIRE from SPIRES. For instance, most of the European countries have a higher usage fraction in INSPIRE, which means means they have generally switched from SPIRES more than the United States, which has more of a half and half distribution. China has a large SPIRES usage metric, which could be due to the Chinese robots mentioned earlier that perhaps were not detected. In the future more of these can be identified and possibly removed. The “Both” columns can give an estimate of proportion is still in the transitional phase, as well as where the usage of those in this transitional phase is concentrated. This does not tell us about the users in the transitional stage, but it does ask questions about how the transitional process works. Do users gradually change, preferring INSPIRE for some tasks and SPIRES for others? Do they have a comparison phase, where they do similar searches on both to see the differences? These kind of questions could potentially be answered by examining in detail the sessions of the users that fall into the “Both” category.

The frustration event data in Table 2 indicates that INSPIRE has a higher occurrence of frustration events than SPIRES. This is because INSPIRE, as a newer service, is still being learned by its users. The data in Table 3 is more interesting, because it shows the kind of problems that cause the frustration events. The high occurrence of frustration events caused by author search keywords is due to misspellings in author names, something that is common in both SPIRES and INSPIRE. It can also show situations where a particular author is expected to appear in the search results but does not. The high occurrence of the journal search keyword in INSPIRE as opposed to SPIRES is due to a flaw in the code of search engine parsing that was present over the time period the data was taken from. This shows the usefulness of the frustration detection analysis in highlighting important issues for the developers. The frustration

detection has already identified an important feature that is broken in INSPIRE, and will continue to prove a valuable tool. In the near future, the algorithm for frustration detection will be made a part of the INSPIRE web statistics module to continually watch for frustration events.

ACKNOWLEDGEMENTS

This work was supported by the Science Undergraduate Laboratory Internship program at SLAC National Accelerator Laboratory and the Department of Energy. I would like to thank my mentor, Travis Brooks, for his leadership in this project. I would also like to thank Joseph Blaylock, Valkyrie Savage, and Mike Sullivan for their assistance.