Background information on the INSPIRE website

[Jun 2013]
Purpose of this document: to provide detailed background information that inform about the specific setup, usage and community specific aspects of this website.
Provided by INSPIRE team.

Table of Contents

1. INSPIRE today

INSPIRE is the main global community information system in High Energy Physics (HEP). This central entry portal to the HEP scholarly information landscape is regularly – often daily – used by basically all HEP scientists. It is the main tool to find and retrieve papers, their references and citations, linking to all the main publishing outlets in HEP. In addition, INSPIRE offers information about authors and their institutions, conferences, HEP experiments and job offers.

The HEP community comprises many sub-disciplines. Overall, we distinguish theoretical and experimental physicists. The partition is almost 50:50 in terms of researchers. They vary in their individual research workflows; e.g. experimental research is being conducted in international (and often large) collaborations. Big experiments are conducted in major hubs, laboratories like CERN. Researchers travel globally to work on such experiments. On the other hand, theorists are rather distributed at often smaller institutions, do collaborate remotely globally, and tend to publish more: about 90% of the articles published in the field. Most likely they are the largest user group of INSPIRE. At the same time, experimentalists very much depend upon the published body of physics research and their collaborations author some of the most read articles in the field.

Articles are published in a preliminary format as preprints, and in final format in scholarly journals run by commercial publishers or scholarly societies. Researchers (and their institutions, funders and policy makers) pay a lot of attention to the individual's research output. The impact on the research community is primarily measured by the number of publications and citations thereof, leading to career and funding decisions. Publication and citations are often dubbed the current "currency" in science.

INSPIRE today holds 1 million metadata records: mostly inherited from SPIRES, its predecessor system, and cumulated over 40 years of activity. These metadata records provide information about virtually all HEP research articles available, and also provide access to these articles, e.g. via links (DOIs, document IDs, etc.) to the respective external sites. INSPIRE indexes and in some cases hosts and directly serves also 300'000 full-text documents (attached to the metadata records) from a partner repository, arXiv, and additional sources. This means that visits to INSPIRE are sometimes rather short, as researchers often look for a specific paper or author for example. This is desirable as an indicator of success: researchers found the information they needed and followed a link to read the relevant article from the original source. INSPIRE has an average traffic of over 2 searches per second from the 50'000-strong HEP research community. A survey conducted in 2007 revealed that HEP scientists predominantly use community-run tools to access information. The two closely collaborating community services INSPIRE (and its predecessor SPIRES) and the eprint repository arXiv.org are the main entry points to scientific information in the field, aggregating and presenting all sources, a sort of benign "monopoly".

INSPIRE is true to its community roots and maintains open lines of communication with its user community. The site provides both contact e-mail addresses and forms, for user requests ranging from general feedback, to author related matters, suggestions of articles or reference corrections.. These requests come in frequently and are recorded in

an RT ticketing system. The tickets are handled as soon as possible by the INSPIRE staff members. To date, the main interaction with the user occurs via email.

Among the most popular user requests are those asking for a correction to a cited reference. These are typically of the form "Paper x cites my paper y, but the citation doesn't appear in INSPIRE, can you fix it?" INSPIRE uses text mining to extract references from pdf files, which works well in most cases but inevitably misses out on some references. The author community is motivated to make sure this inter-article linking works, because the links are directly reflected in their citation counts, and ultimately feed in their evaluation for promotions, accounting of research funds, or the overall academic job market.

Another common request involves the challenge of author disambiguation. For purposes of searching and compiling publication lists, INSPIRE uses advanced algorithms to try to deduce if "Richard Feynman" is the same person indicated on an article as "R Feynman", "RP Feynman", and so on. This is difficult to do with high accuracy, especially in the case of common family names. Readers often write to report that papers on their automatically generated "author profile" page do not actually belong to them, or that ones they did author are missing.

More recent additions to our community engagement features are tweets (presented as "news" on the INSPIRE front page) and a blog. Those reveal popular to inform users about new features or warn them in the rare case of brief outages in a feature.

2. History, development and community specific aspects

History and development

INSPIRE is based on a strong history of community tools. Starting in the early 1960's, HEP researchers adopted a simple solution to share their results faster than the traditional peer-review and journal distribution system: mailing their research, in its initial *preprint* form, to their colleagues before sending it to a journal. The advantages of speed and the sense of community vastly overweighed the risk of spoofing. Research libraries such as CERN/Geneva became hubs, started building metadata records of those preprints at par with other material and further contributed in making this research searchable and retrievable, and further re-distributed information. Automation in the late 1960s allowed the SLAC/Stanford library to create a computerized system, SPIRES, as a niche of the ecosystem of communication where all unpublished preprints would be recorded. A few years later, the library at DESY/Hamburg joined, contributing mainly metadata on journal articles and conference papers as well as keywords from a HEP thesaurus. Fermilab/Chicago became the 3rd partner. In December 1991 SPIRES became the first database on the web and the first website outside of Europe.
Also in 1991 a repository for preprints was developed, the arXiv (http://arxiv.org/). Distribution became easier: HEP researchers directly submit their papers to disseminate them and receive feedback as early as possible. This so called eprint repository changed

HEP scholarly communication forever – researchers first share their research here and then submit to the traditional journals. Its crucial role in the research process is manifested by the fact that HEP scientists predominantly read articles on arXiv, very rarely visit journal websites and regularly cite arXiv articles before their publication in a journal.
From the start, SPIRES collaborated closely with arXiv, daily ingesting records relevant to HEP. SPIRES added bibliographic value for users (affiliations, references, citations, links to published versions, conference information and more) and acted as the entry point for wider literature searches. The two information systems played complementary roles: arXiv became the place to deposit new material, SPIRES the place to perform searches.
The HEP repository ecosystem has evolved fast: CERN has joined and SPIRES has been superseded by INSPIRE, built on the Invenio Open Source Digital Library software (http://invenio-software.org/). In October 2011 INSPIRE went into full production (after a year of beta release). In addition to the core research literature database, specialised databases are now an integral part of the INSPIRE offering, including HEP Conferences, HEPNames (a directory for HEP researchers), HEP Institutions, HEP Jobs, and HEP Experiments.

Today, INSPIRE maintains complex interconnections with other HEP community resources. For example:

- INSPIRE daily receives an automated feed from arXiv. Metadata and full-text are indexed for searching. Requests for PDFs are redirected back to arXiv.
- INSPIRE crawls the output of publishers of HEP journals, or directly receives periodic feeds, and makes this data available to the HEP community, uniting records of published instances of articles with their pre-print versions wherever possible.

Beyond this initial core of arXiv preprints and publisher feeds, other community-based services that have independently evolved in this ecosystem have been integrated carefully into the INSPIRE database. For example the HEPData service (http://durpdg.dur.ac.uk/, Durham UK) is an *ante litteram* data repository in the field. Working closely with Durham, INSPIRE records now link to HEPData datasets.
Individual HEP laboratories are also independent nodes in the ecosystem, each producing specific, institutional, information fed back to INSPIRE when relevant to a wider community. INSPIRE locates, harvests, and indexes such content together with the other literature in the field. These are presented in the respective authors page of INSPIRE, aggregating all the scientific output of members of the community.

Some specific features

References and citations

On a daily basis, content is automatically ingested into INSPIRE and made available to the user. The metadata is curated manually, so that it is consistent and up-to-date. A major concern to data curators and users are correct citations to articles on INSPIRE. Many feedback tickets request corrections of reference lists or contain complaints about

missing updates (as the backlog of reference curation is currently several months). This is important to note as references and thus citations of papers are, as mentioned before, the "currency" in science. Since the increasing curation workflow surpasses the capacity of the INSPIRE cataloguers INSPIRE tries more and more to engage users in curating records themselves, thus improving the metadata quality and speeding up curation. One of the first crowdsourcing applications was the introduction of user webforms for reference corrections.

Plot extraction

Plots are extracted from the LaTeX sources of arXiv papers and displayed in the detailed format of INSPIRE records. A script to extract plots from pdf files is currently under development.

Author pages

The INSPIRE author pages (see for example the record for one of the most prolific scientists active today at www.inspirehep.net/author/J.Ellis.1/) present a unique view of a researcher's profile, including current and previous affiliations, collaborators, paper counts, keywords, citation counts, and biographical information (education, thesis advisor). These pages are created automatically from the metadata present in each article record, as well as integrating information spontaneously contributed by users to the HEPNames directory. These pages now provide a standardized view of the contributions of each researcher.

As mentioned above, these author profile pages depend upon automated routines that provide author name disambiguation. Since these routines can never be perfect, the author profile pages contain a link ("This is me…") to an interface where researchers can unambiguously verify their publications list on INSPIRE ("paper claiming"). This insures that the publications are assigned to their personal name persistently. In the past two years the INSPIRE team has invited researchers via email to use this tool and correct their publication profiles on INSPIRE. Within the first year, response rates were rather high, on average about 40%.

Plans are being made to build on this successful experience and upgrade the crowdsourcing activities. Ideally, we would like to provide authors and readers with better tools and a more intuitive interface with paper claiming. We may also wish to extend the data that is available for crowdsourced input. This is part of an overall initiative within INSPIRE to further the crowdsourcing opportunities.

This overview of all the materials that are connected to a HEP researcher has grown in the past years. Thus, the author page is an assemblage of the individual parts on this site, which have been put together through time (the citation counts, the affiliation history, the HEPNames directory, the academic history). As more material is published, beyond text (e.g. snippets of code or some scientific data) it is expected that the complexity of this page will grow. Moreover, users request advanced metrics combining citations in different ways (already now a second page with more metrics is accessible from the author pages, e.g. http://inspirehep.net/author/M.Weber.1/; and http://inspirehep.net/search?ln=en&p=author%3A%22M.Weber.1%22&of=hcs2).

Thus, we are facing challenges to the information architecture, to the display of data on author pages as well as to user navigation.

Search syntax and services

On INSPIRE one can use three different search syntax:

- Traditional SPIRES syntax: a legacy of the previous database system dating from the '70s. Because of its short-hand notation this syntax is very efficient and therefore still very popular with 80% of the users still using it. (http://inspirehep.net/help/search-tips)
- Invenio search syntax: made possible as of 2010 from the new underlying software with second-order and third-order operators such as refersto:citedby:recid:12345 [give me the articles referring to the articles cited by the record 12345, see also http://inspirehep.net/help/invenio-search-tips
- Google like free-text search: searches in all indexes (author, abstract, title, keywords), possible since 2010.

INSPIRE provides several specific search interfaces.
For fulltexts hosted on INSPIRE phrase searching is offered.
To facilitate the inclusion of papers within CVs, bibliographies etc, search results are offered in several output/export formats, including LaTeX variants. These are extremely important, as HEP researchers use INSPIRE to generate reference lists to be included in their own papers.

3. Usage of the INSPIRE website – A brief overview

The following analysis is based on a 2-weeks dataset in March 2013, from March 5th to March 18th). The software PIWIK has been used to take this data snapshot. Access through robots and RSS have been excluded. The statistics and plots presented are based on the data gathered through PIWIK.
A list of the most popular searches is given in Fig. 6. It highlights the main tasks of the users on INSPIRE.

The users are predominantly from the US and Europe (Fig. 1; Switzerland is rather high due to CERN usage) and reflect each country's contribution to the literature of the field, through the number of active scholars in this discipline. The results are reflected in the browser language choice as well. It has to be noted that the main working language in HEP is English by far.

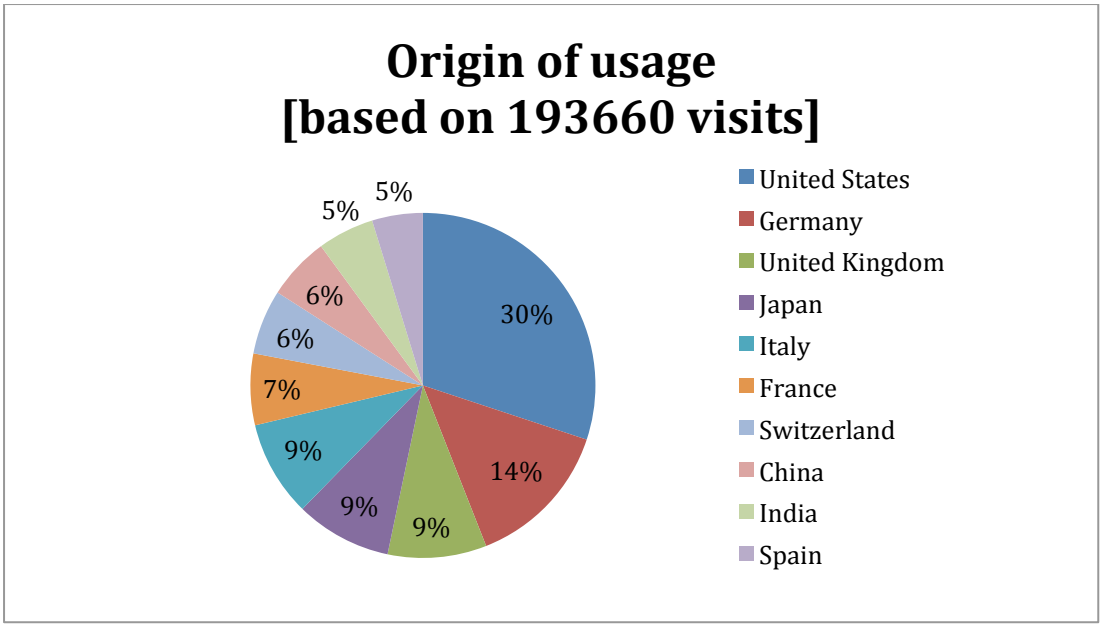

Fig. 1
Origin of usage (%), Top 10 countries only. Usage from CERN is counted as Switzerland. In total: 193660 visits counted

Users predominantly use Safari and Firefox, followed by Chrome, as browser to access the INSPIRE website. Internet Explorer only makes up a small fraction of the usage (Fig. 2).
Only 4% of the usage comes from a mobile device (not shown in this report). The majority of the usage is recorded from desktop applications.

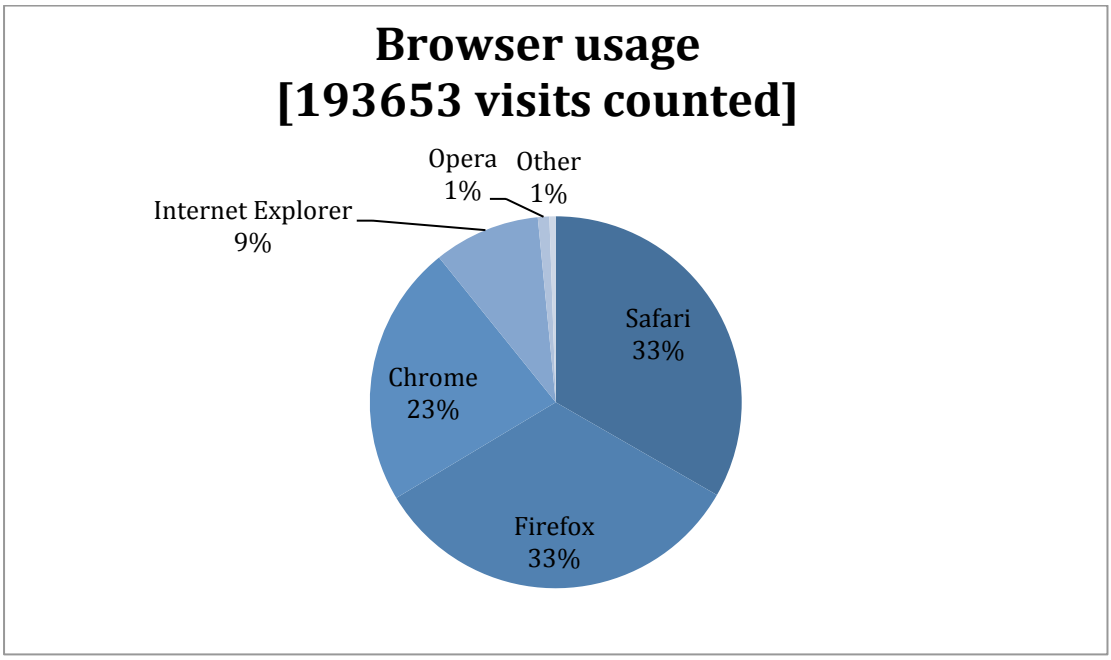

Fig. 2
Browsers used to access INSPIRE, in total: 193653 visits counted.

The main usage is observed during the mid-week (Fig. 3), but as the cycle of science goes, it only halves over the weekend. The frequent usage of INSPIRE over the weekend is also reflected in the feedback tickets that are received on INSPIRE. As an example: a glitch in the citation counts that happened on a Saturday evening CET resulted in an almost immediate email storm to complain about the individual researchers' citations counts on the author pages.

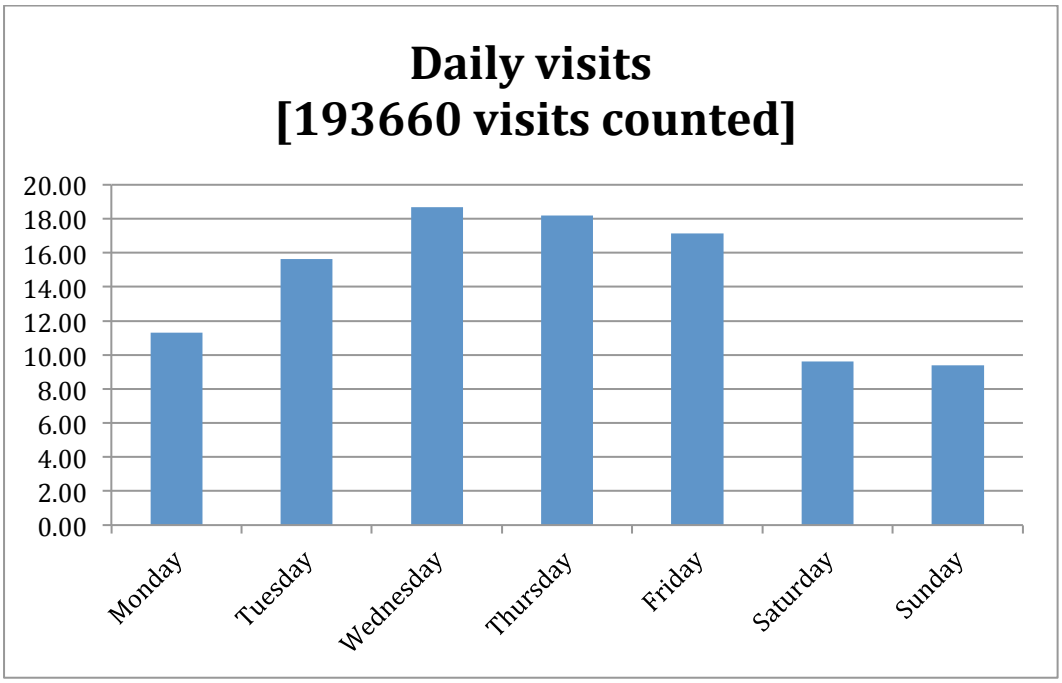

Fig. 3
Visits over week in %, in total: 193660 counted.

4. "Known issues": Results of a user feedback survey on INSPIRE

A brief feedback survey was conducted in March 2013.
Within one week, the survey yielded 560 responses. The respondents can be described as follows: The vast majority of the respondents can be considered power users as they use INSPIRE on a daily basis. This is not surprising as the survey was only live for a week, so it is biased towards daily users. Given that 51% have more than 10 years of experience with SPIRES and INSPIRE it can be assumed that they are rather senior researchers (Fig. 4 below). It remains an open question how representative this sample is of the entire user community of INSPIRE, however. Such potential biases are taken into account in the following analysis.

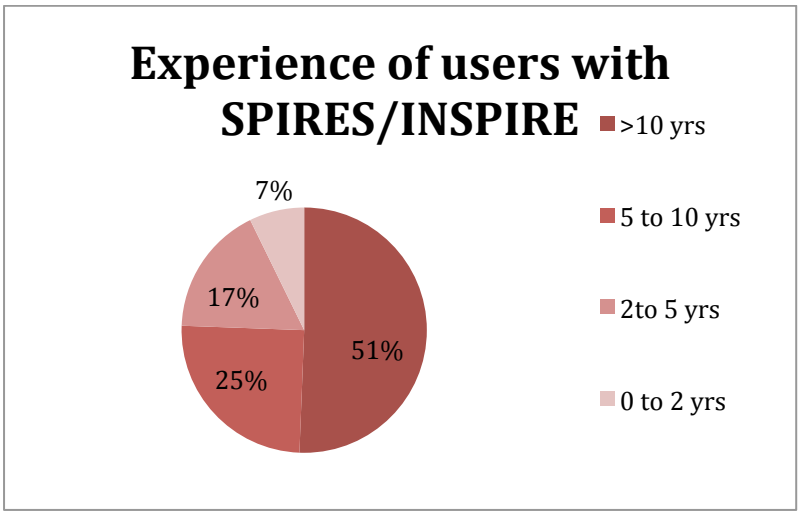

Fig. 4

Seniority and experience of INSPIRE users

Survey respondents were asked about their main field of research. The majority of respondents are from the core HEP fields (more than 80%), but neighbouring fields are also covered. This is important to note as they might be looking for publications of adjacent fields that are not (fully) covered by the HEP core content.

Reported issues "What do you like the least on INSPIRE, and why?":

Issues are divided in two main categories, those related to the use of INSPIRE, and those related to "customer service", intended as the preparation of content before display and the speed/efficiency of reacting to suggested corrections of the content. The key issues in the first category (see Fig. 5), relevant for this study are in particular these two aspects:

- Search functionalities: The search syntax is the core of INSPIRE. When using the SPIRES/INSPIRE syntax, putting the correct search queries together is crucial to get the correct results. The free text responses highlight that researchers struggled with the different search syntax on INSPIRE. This has also been seen in

the log files (see chapter 2). One can clearly identify how researchers struggle to find the correct results in the individual search session. They need several search queries to retrieve the correct search result page. They are looking for a specific publication in a journal and cannot find it straight away; so they type the search slightly different each time. The level of frustration can easily be seen in such sessions.

- Author related issues: refers to the overall increased interest in person centric services. Users want correct publication and citation counts on their personal author page and their data up to date. Specifically, the mix up of common author names (author disambiguation) and the assignment of papers to the wrong person were criticized..

Author-centric concerns also play often a role in comments about directories/databases not being up to date or not easy to use (especially HEPNames) and in comments regarding customer service. The most frequent complaints about customer service refer to the often long time delay in the curation und update of references and citations.

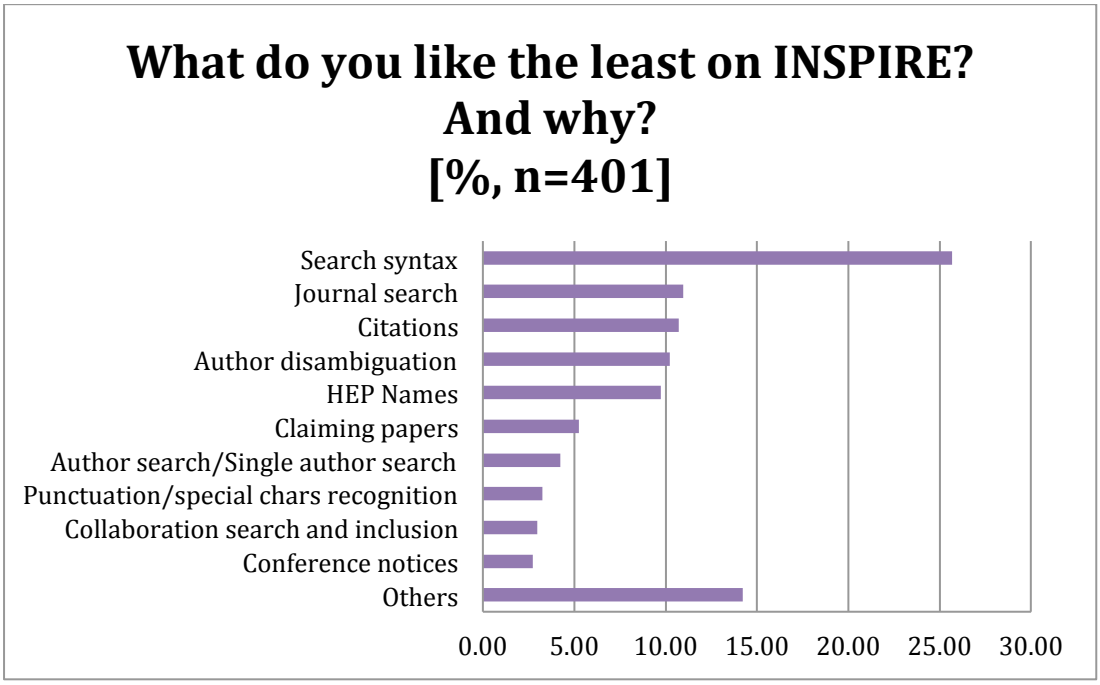

Fig. 5

Reported issues on INSPIRE. For this plot only the tags relevant to UI/UX are shown (complaints about INSPIRE services have been excluded here). The tags can be described as follows:

- *search syntax: is complicated, counter-intuitive, outdated*
- *Journal search: is not flexible and user-friendly*
- *Author disambiguation: author names are mixed up, publications are not assigned to the correct author (page)*
- *Citations: Unlisted citations of publications, stats give incorrect number of citations*
- *HEPNames: Using and updating HEPNames (a directory) is complicated, data is sometimes outdated*

- *Claiming papers: workflow for researchers to get their papers assigned to their profiles is clumsy and complicated in case of common names*
- *Author search: difficult when looking for single authored papers etc.*
- *Special chars recognition: are not recognized, there is no tolerance to spelling mistakes or punctuation (and no suggestions given) which makes search harder*
- *Collaboration search: difficult to distinguish publications published by large collaborations from the ones authored by a single person in the search results output*

Aspects the users like: "What do you like the most on INSPIRE, and why?"

INSPIRE users like the speed of the INSPIRE website. This is not surprising as the majority of the respondents have a long experience of SPIRES usage (more than 10years) and the old system had become notoriously slow.

The survey results (Fig. 6) furthermore underline an impression gained from the daily feedback emails received by the INSPIRE team: researchers pay a lot of attention to the correct references and their citations. INSPIRE lists their citations on the author pages. In spite of criticizing some noticeable deficits in reference handling and author profiles users very much appreciate the advancement from SPIRES to INSPIRE with regard to their personal profiles.

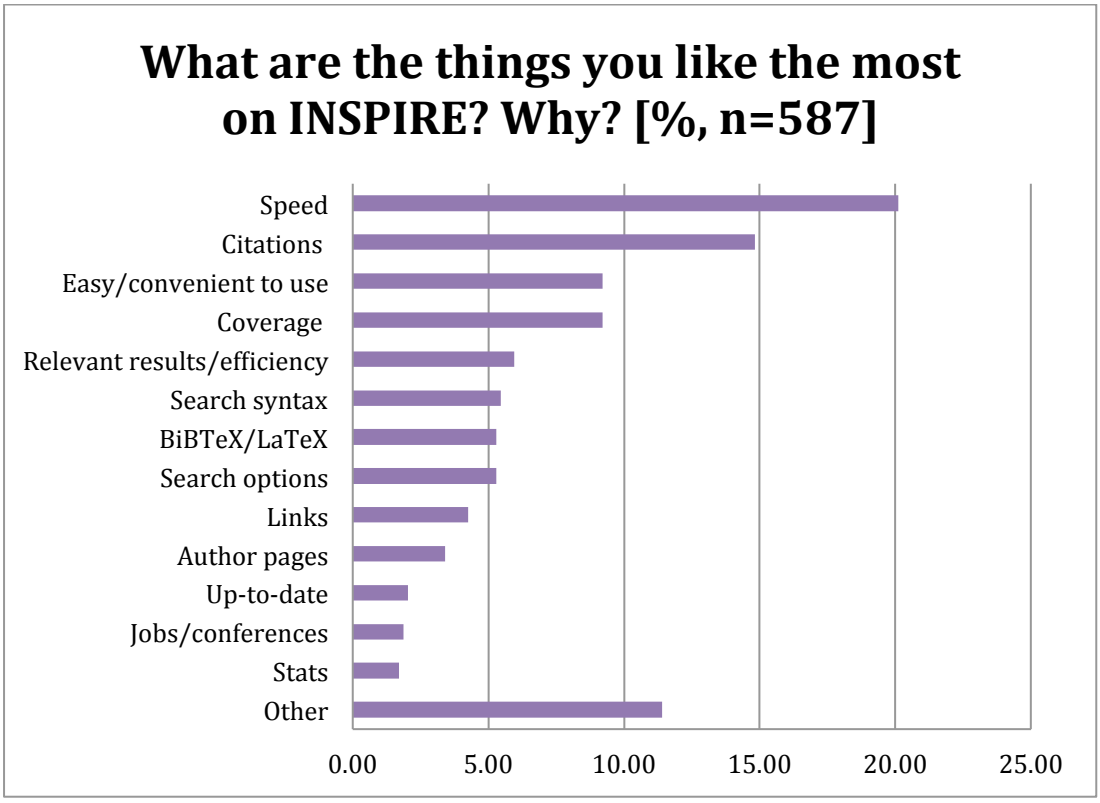

Fig. 6

Users replies to the question "What are the things you like the most on INSPIRE? Why?". Please note that one user might have given longer responses and so several tags have been

assigned to his reply. That's why the number of tags exceeds the number of overall responses to the survey/question. The main tags can be described as follows:

- *Speed: INSPIRE is much faster than its predecessor SPIRES*
- *Citations: Better and more complete reference and citation counts, and citations summary of individuals*
- *Coverage: All the relevant content for HEP is there*
- *Easy/convenient to use*
- *Relevant results: They are reliable and accurate and the content matters to HEP*
- *Search options: are flexible enough to obtain the correct results*
- *BibTeX/LaTeX output: the community uses these tools to write publications and to compile the reference lists so the export of metadata in such formats is convenient.*
- *Links: INSPIRE provides outlinks to arXiv or other sources*
- *Search syntax: powerful, built on the traditional and good SPIRES syntax*
- *Author pages: relevant information about the author, easy to find info about author*
- *Up-to-date: has the latest HEP relevant publications*
- *References: good display and tracking of references*
- *Jobs/conferences: useful tool to announce/find jobs and conferences*
- *Stats/Metrics: useful information on the research output of individuals and collaborations*