INSPIRE: a new scientific information system for HEP

R Ivanov and L Raae

European Organization for Nuclear Research, CERN CH-1211, Geneva 23, Switzerland

E-mail: Radoslav.Ivanov@cern.ch, Lars.Christian.Raae@cern.ch

Abstract. The status of high-energy physics (HEP) information systems has been jointly analyzed by the libraries of CERN, DESY, Fermilab and SLAC. As a result, the four laboratories have started the INSPIRE project – a new platform built by moving the successful SPIRES features and content, curated at DESY, Fermilab and SLAC, into the open-source CDS Invenio digital library software that was developed at CERN. INSPIRE will integrate current acquisition workflows and databases to host the entire body of the HEP literature (about one million records), aiming to become the reference HEP scientific information platform worldwide. It will provide users with fast access to full text journal articles and preprints, but also material such as conference slides and multimedia. INSPIRE will empower scientists with new tools to discover and access the results most relevant to their research, enable novel text- and data-mining applications, and deploy new metrics to assess the impact of articles and authors. In addition, it will introduce the "Web 2.0" paradigm of user-enriched content in the domain of sciences, with community-based approaches to scientific publishing. INSPIRE represents a natural evolution of scholarly communication built on successful community-based information systems, and it provides a vision for information management in other fields of science. Inspired by the needs of HEP, we hope that the INSPIRE project will be inspiring for other communities.

1. Introduction

In late spring 2007 four high-energy physics (HEP) laboratories, The European Organization for Nuclear Research (CERN), the Deutsches Elektronen Synchrotron (DESY), the Fermi National Accelerator Laboratory (FNAL) and the Stanford Linear Accelerator Center (SLAC), ran a user poll to analyze the current state of HEP information systems. The goal was to achieve a better understanding of the perceptions, behaviors and wishes of the end users of these information systems. The poll received more than 2100 answers, representing about 10% of the active HEP community worldwide.

The poll showed that community-based services dominate this field of research with the metadata-only search engine SPIRES-HEP [1] being the primary information gateway for most scholars. Users also gave their preferences regarding existing functionalities like access to full text and to citation information, and a list of features that they would like to have in the coming years. The results showed that the scholars attach paramount importance to three axes of excellence: access to full text, depth of coverage and quality of content [2].

Based on the results of the poll representatives from the four labs decided to investigate further how a closer collaboration could fully match the community expectations. A feasibility study was conducted and one started experimenting with replicating SPIRES content and features in CDS

Invenio, a digital library software suite developed at CERN. These experiments concluded successfully and in May 2008 the INSPIRE project was announced. This article aims to introduce INSPIRE and some of the platform's key features.

2. SPIRES and CDS Invenio

2.1. SPIRES

The SPIRES-HEP database stores bibliographic information about the literature in the field of High Energy Physics. SPIRES-HEP was born in 1974 and was based on SPIRES DBMS, using an IBM mainframe and command line interface. In the 1980s, an email interface was added and in the early 1990s, the first US Web Server was established at SLAC to provide access to the SPIRES-HEP database [3].

Today the service is being run by SLAC, DESY and Fermilab and is providing high quality metadata with human-proofed publication information, links to full text, author affiliations and much more. The before mentioned poll showed that SPIRES is the most popular information system in the HEP community, with 48.2% replying that it is the system they use the most [2].

Nevertheless being such a veteran system, SPIRES now suffers from its aging technology (SPIRES DBMS), resulting in scalability and maintenance issues. Therefore in the recent years one has been searching for a new platform to host the content of SPIRES [4].

2.2. CDS Invenio

CDS Invenio [5], developed and maintained at CERN, is a suite of applications which provides the framework and tools for building and managing an autonomous digital library server. The software is licensed under the terms of the GNU General Public License (GPL), and covers all aspects of digital library management.

CDS Invenio is designed to support moderate to large size (> 1 million records) systems, while maintaining very fast search speeds. As an example, the CERN library CDS Invenio installation contains 900 000 records, yet the search for "'lepton" (16 433 hits) takes 0.18 seconds, and displaying the first page of results to the user in browser typically less than 1 second.

CDS Invenio relies on acknowledged standards such as MAchine-Readable Cataloging (MARC) for storing bibliographic data [7] and Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) [8] for the exchange or harvesting of metadata from external systems. The system enjoys the support of a dedicated development team at CERN, while also receiving contributions from external contributors in its user community at about 25 different institutions worldwide.

3. INSPIRE

The high quality metadata in SPIRES, combined with the fast and scalable software of CDS Invenio, has seemed like a good match to meet the user expectations that were expressed in the poll. The laboratories of CERN, DESY, Fermilab and SLAC agreed to build INSPIRE, a new global HEP information platform, by (i) merging the content of the SPIRES and CDS databases, (ii) using the CDS Invenio software for searching and displaying the records and (iii) reproducing and extending the functionality of SPIRES by new CDS Invenio modules. However, it was clear that this would have to be done without causing disturbance to the users, meaning that the preservation of SPIRES features, interface and syntax already familiar to the user, would have to be the first step.

After reproducing SPIRES features, the INSPIRE collaboration is now focusing on cataloguer-level functionalities. The objective is to build strong native tools to enable libraries from the four institutes to share the workload of data input and verification.

The main benefits of INSPIRE can thus be listed as follows.

- For users: fast search, access to full text, high-quality metadata, new bibliographical metrics (including citation analysis) and the possibility to contribute records or suggest corrections to the metadata.

- For the participating institutes: tools for editing and checking the metadata, and better software-assisted coordination between the institutes.

These features will be discussed in more details below.

Figure 1. INSPIRE is based on CDS Invenio software and SPIRES content.

4. Powerful search capabilities

The CDS Invenio software provides INSPIRE with a powerful search engine. Using indexes designed specifically for rapid access most queries are executed in milliseconds, even in a repository of more than one million records.

User adaption is strengthened by supporting both SPIRES and "Google-like" search syntax. Simple and advanced search interfaces are available to meet the requirements of different user groups. Search results can be sorted by different criteria and are available in several output formats. In addition users can choose various ranking mechanisms for the search engine to apply when displaying the results.

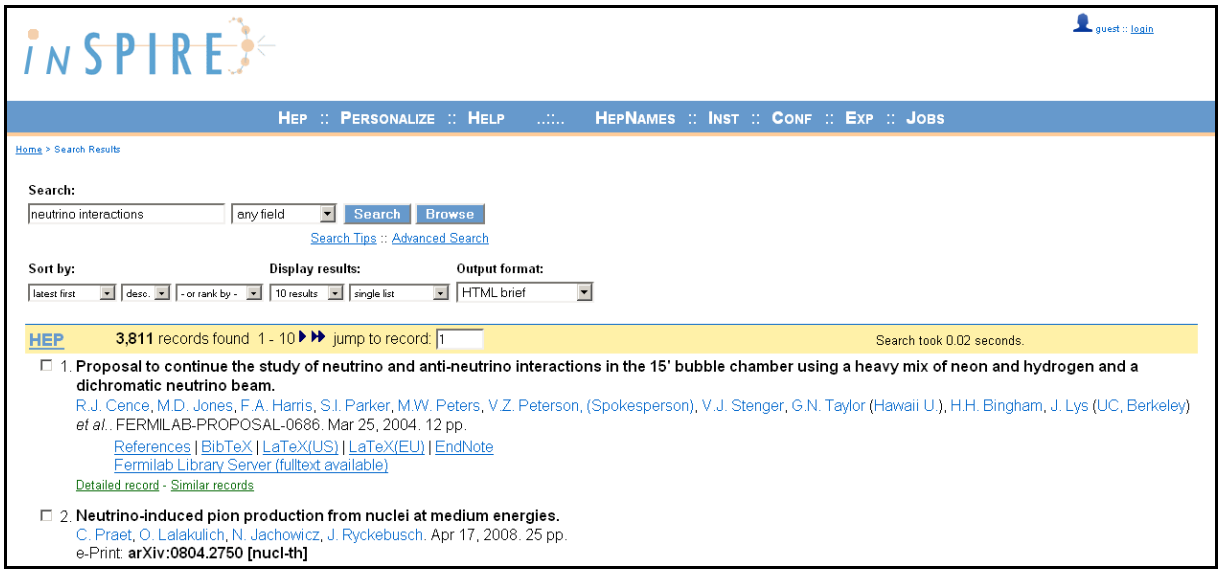

Figure 2. INSPIRE search interface.

5. Citation analysis

One way in which INSPIRE will bring additional value to the HEP community is by doing citation analysis of the articles in the repository. Analysis of the pairs "citing article - cited article" in INSPIRE's database, lets us generate a summary about the citations of every paper. Thus, we can see citation summaries for an author, an author's institute (using author affiliation data), year of publication etc.

Figure 3 shows the citation summary for an article. It contains information and statistics about the records citing this article and the records co-cited with the article, and a diagram of how the citations are distributed over time.

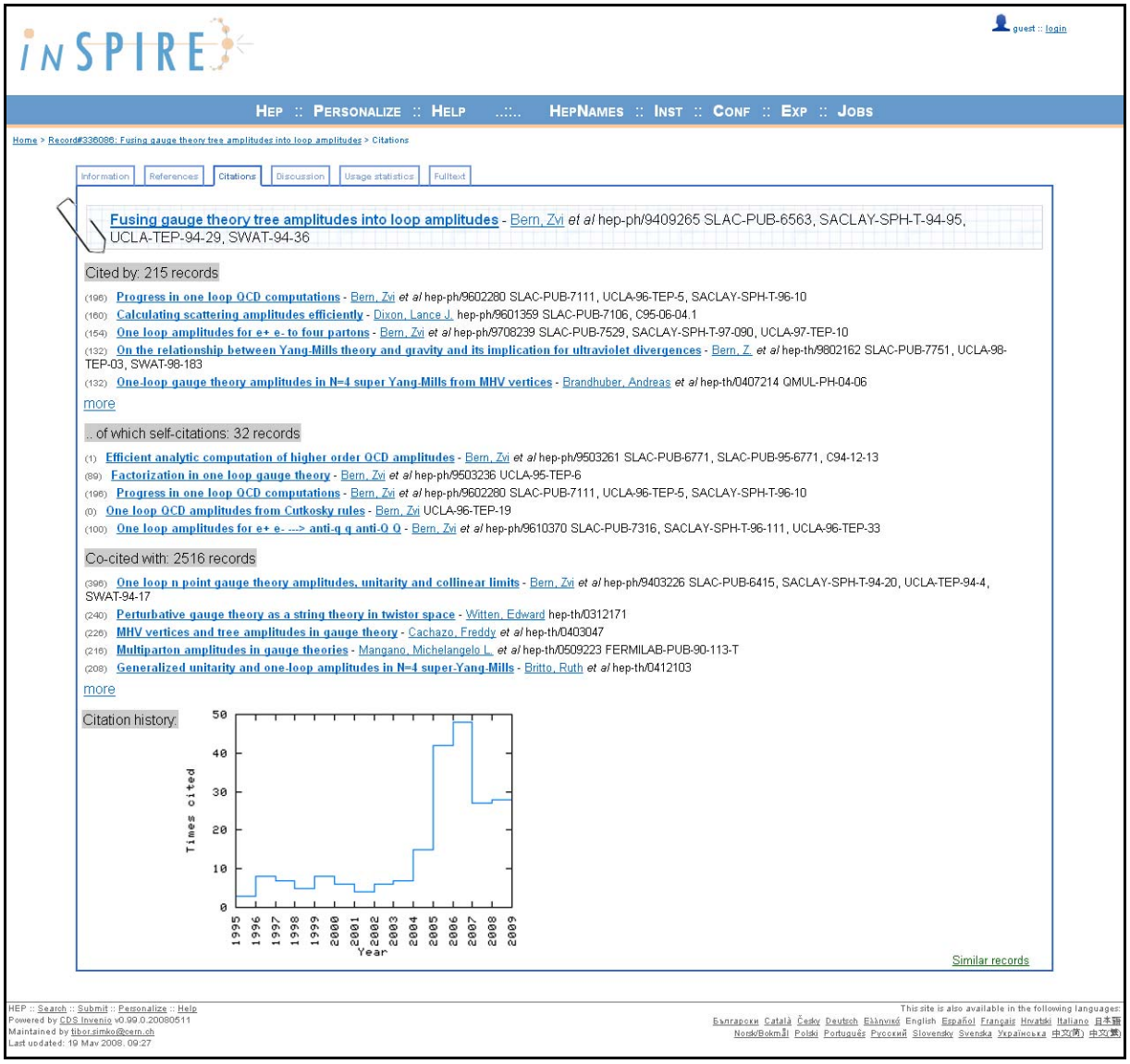

Figure 3. Citation summary page.

This is only a sample of the potential of INSPIRE to deploy new metrics and perform analysis based on available data. It might be interesting to know that analysis of the current data in INSPIRE shows that an average article cites 12 other articles in the database, and the most cited paper in the database is "A Model of Leptons".

6. Author summary

Author summary pages, like the one shown in figure 4, are another recent addition to INSPIRE. They allow assessment of author activities summarizing what is known about each author in INSPIRE's database. This includes the author's home institute (or many of them since authors often work in different sites during their career), most popular keywords used in the author's articles, most frequent co-authors, a breakdown of articles based on their type (e.g., books, conference presentations, lectures) and a breakdown of articles based on their citation data.

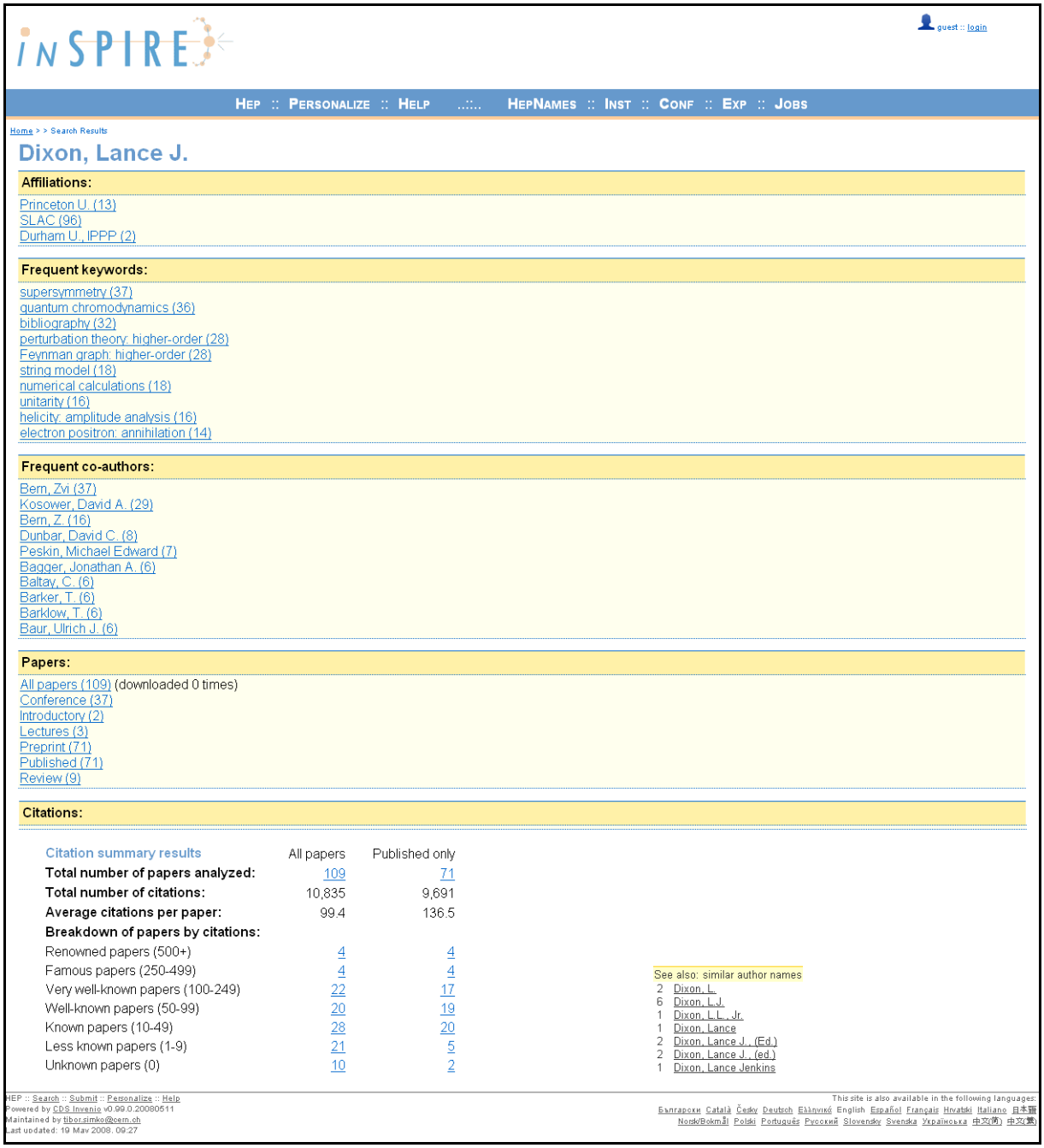

Figure 4. Author summary page.

7. Back-office tools

In INSPIRE, most records will come into existence either through manual inputting or through harvesting of records from external sources. All the records are subject to preliminary processing before they are accepted. Catalogers from the collaborating laboratories have the important job of controlling, cleaning and enriching the flood of data which will be going into the system [6].

In order to streamline the workflow and ensure delivery of high quality content, the INSPIRE collaboration is building powerful tools for data processing. A variety of modules and tools will play different roles in order to meet the following objectives:

- enable global cooperation between catalogers at the different labs by building an optimized and distributed cataloguing environment
- increase metadata quality without generating significant overhead
- automate as much of the catalogers work as possible
- provide tools to let the catalogers do their job easily, efficiently and without needlessly repetitive or pointless tasks

While many tools are still in the process of development, some are already production ready. This includes tools for automatic extraction of keywords and references, metadata editing via Web interface and automated testing of metadata for compliance with quality standards and existing knowledge bases. These knowledge bases will contain authoritative files of authors, institutions and other important datasets.

The automated tools will do as much work on the records as possible before they reach the cataloguer. A graphical user interface for interactive record editing will gather the results from the other modules and present them to the cataloguer. The objective is to keep the common cases as efficient as possible and the actions that have to be performed manually by the cataloguer to an absolute minimum.

All the developments are done as direct enhancements to the CDS Invenio software, since most of them are not specific to INSPIRE and are applicable to a general cataloguing workflow. This will allow all systems that are using CDS Invenio software to benefit from the innovative developments related to the INSPIRE project.

8. User personalization and collaborative tools

A stated objective of INSPIRE is that it will be a community-based and user-driven information platform. By making use of the Web 2.0 philosophy of utilizing community resources to enrich data, INSPIRE will introduce new features to facilitate user collaboration and information sharing. One such feature is user-tagged content, meaning that users are giving the opportunity to assign tags or keywords to a document. When asked about their willingness to participate in this kind of volunteer work, 63% of the respondents said that they were willing to spend between five minutes a day and an hour a week.

User personalization is another feature of contemporary web systems. CDS Invenio already supports user-defined baskets of documents and automated e-mail alerts, and it has a multilingual interface supporting 20 languages, so most users can choose to use the system in their own language. However, such tools require user authentication, for which a coherent solution for all users should be addressed by the INSPIRE partners.

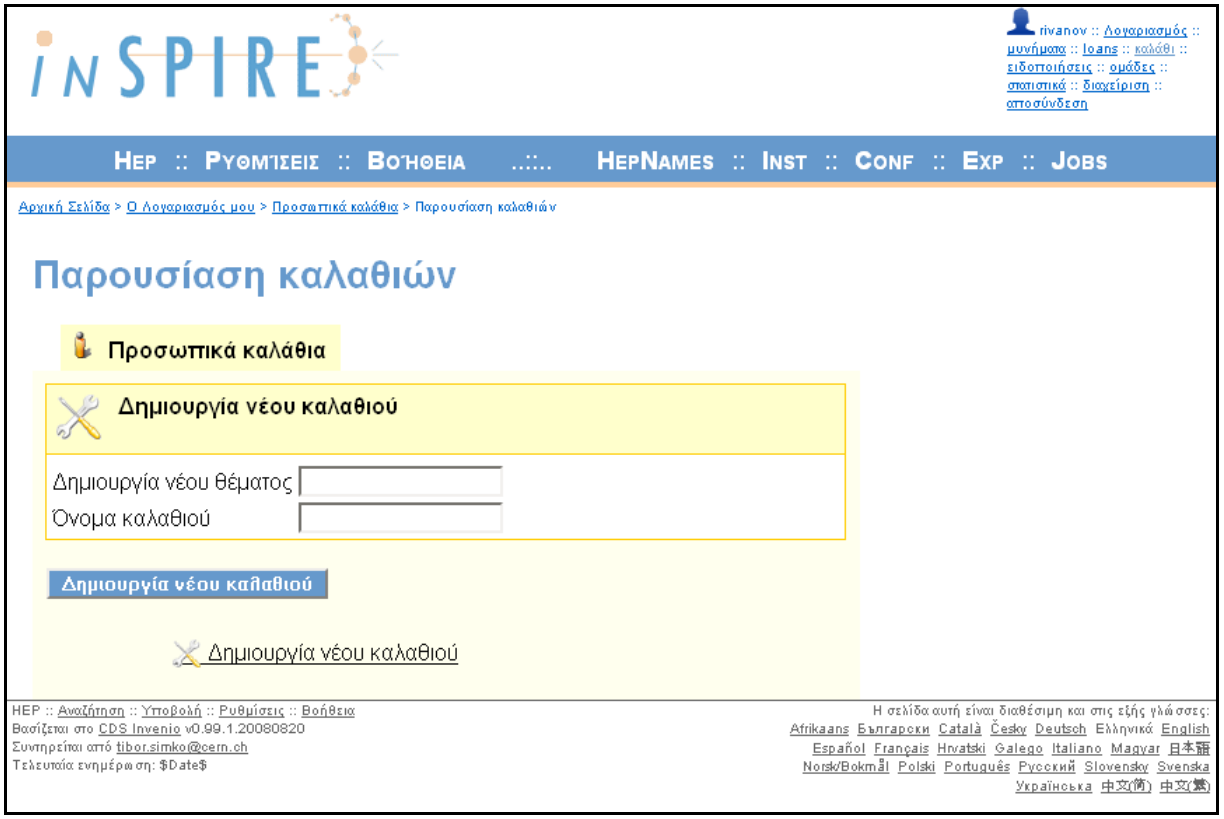

Figure 5. INSPIRE user interface in Greek.

9. Conclusion

INSPIRE is an innovative platform resulting from the efforts of CERN, DESY, Fermilab and SLAC to combine the features and content of SPIRES, one of the most popular HEP information systems, with the free, open-source digital library software CDS Invenio.

The system will reproduce the currently existing features of SPIRES and will also introduce new tools in order to meet the growing needs of the scientific community. Citation analysis and author summary pages are only an example about the potential of INSPIRE to perform analysis based on available data. New back-office tools are being developed to support the input and editing workflow of the libraries. User personalization and collaborative tools together with open access to full text articles and other scientific information will support the scientists in the HEP community, and help them to access the results most relevant to their research.

INSPIRE is a new user-driven information system that has the goal to serve the HEP community, providing free access to high quality content, empowering the users with new tools supporting their information needs.

As a project, INSPIRE is a collaboration between four institutions. However, for the output of the project to be successful, INSPIRE will need to inspire the HEP user community to participate and share.

References

[1] http://www.slac.stanford.edu/spires/ [Last visited May 11, 2009]

[2] Gentil-Beccot A, Mele S, Holtkamp A, O'Connell H B and Brooks T C 2008 Information resources in high-energy physics : surveying the present landscape and charting the future course 2008 J. Am. Soc. Inform. Sci. Technol. 60, 150–160

[3] Addis L 2002, http://www.slac.stanford.edu/spires/papers/history.html [Last visited May 11, 2009]

[4] Holtkamp A 2008 Inspire Overview (slides), HEP Information Resource Summit, https://indico.desy.de/materialDisplay.py?contribId=6&sessionId=18&materialId=slides&confId=800 [Last visited May 11, 2009]

[5] http://cdsware.cern.ch [Last visited May 11, 2009]

[6] Raae L 2009 Interactive editing and cataloging interfaces for modern digital library systems, http://cdsweb.cern.ch/record/1174561 [Last visited May 11, 2009]

[7] MARC 21 Format for Bibliographic Data http://www.loc.gov/marc/bibliographic/ [Last visited May 11, 2009]

[8] Open Archives Initiative Protocol for Metadata Harvesting http://www.openarchives.org/OAI/openarchivesprotocol.html [Last visited May 11, 2009]