Journal of Physics: Conference Series **331** (2011) 082005 doi:10.1088/1742-6596/331/8/082005

Physicists Get INSPIREd: INSPIRE Project and Grid Applications

Jukka Klem, Jan Iwaszkiewicz
CERN, CH-1211 Genève 23, Switzerland

E-mail: Jukka.Klem@cern.ch

Abstract. INSPIRE is the new high-energy physics scientific information system developed by CERN, DESY, Fermilab and SLAC. INSPIRE combines the curated and trusted contents of SPIRES database with Invenio digital library technology. INSPIRE contains the entire HEP literature with about one million records and in addition to becoming the reference HEP scientific information platform, it aims to provide new kinds of data mining services and metrics to assess the impact of articles and authors. Grid and cloud computing provide new opportunities to offer better services in areas that require large CPU and storage resources including document Optical Character Recognition (OCR) processing, full-text indexing of articles and improved metrics. D4Science-II is a European project that develops and operates an e-Infrastructure supporting Virtual Research Environments (VREs). It develops an enabling technology (gCube) which implements a mechanism for facilitating the interoperation of its e-Infrastructure with other autonomously running data e-Infrastructures. As a result, this creates the core of an e-Infrastructure ecosystem. INSPIRE is one of the e-Infrastructures participating in D4Science-II project. In the context of the D4Science-II project, the INSPIRE e-Infrastructure makes available some of its resources and services to other members of the resulting ecosystem. Moreover, it benefits from the ecosystem via a dedicated Virtual Organization giving access to an array of resources ranging from computing and storage resources of grid infrastructures to data and services.

1. INSPIRE project

There is a long tradition of distributing preprints of scientific results and operating digital repositories in High-Energy Physics. One of the major services has been the SPIRES database [1] run by the Stanford Linear Accelerator Center (SLAC) since the late 1960's as a database of particle physics literature. SPIRES was built in 1974 based on an IBM mainframe and command line interface. In 1991 it became the first web site in North America and has attracted around 50,000 searches per day from particle physicists around the world. SPIRES has been a very successful system but now suffers from ageing technology resulting in performance and maintenance issues.

INSPIRE [2][3] is the next-generation High Energy Physics information system built by major laboratories CERN, DESY, Fermilab and SLAC. INSPIRE combines the successful SPIRES database content, curated at DESY, Fermilab and SLAC, with Invenio digital library technology developed at CERN. SPIRES brings to INSPIRE trusted and well curated content, experience in managing HEP information resources and a close relationship with the worldwide user community.

INSPIRE supports the SPIRES specific search syntax that many users know very well, and in addition Google-like free text searches can be made in metadata and document fulltext. Invenio has a powerful seach capability and large repositories like INSPIRE can be searched within a fraction of a second. Current INSPIRE seach interface is shown in figure 1.

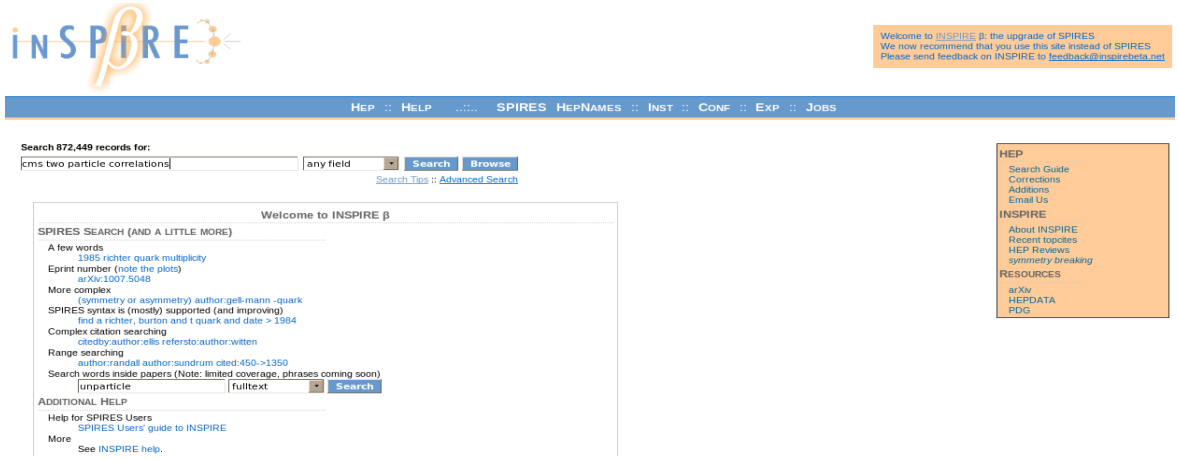

Figure 1: INSPIRE search interface

For each article INSPIRE has a detailed record page that shows article name, authors, publication information, abstract, keywords, links to fulltext documents as well as links to a list of references and citations. For a part of the documents figures are extracted and displayed in the article page. An example of an article page is shown in figure 2.

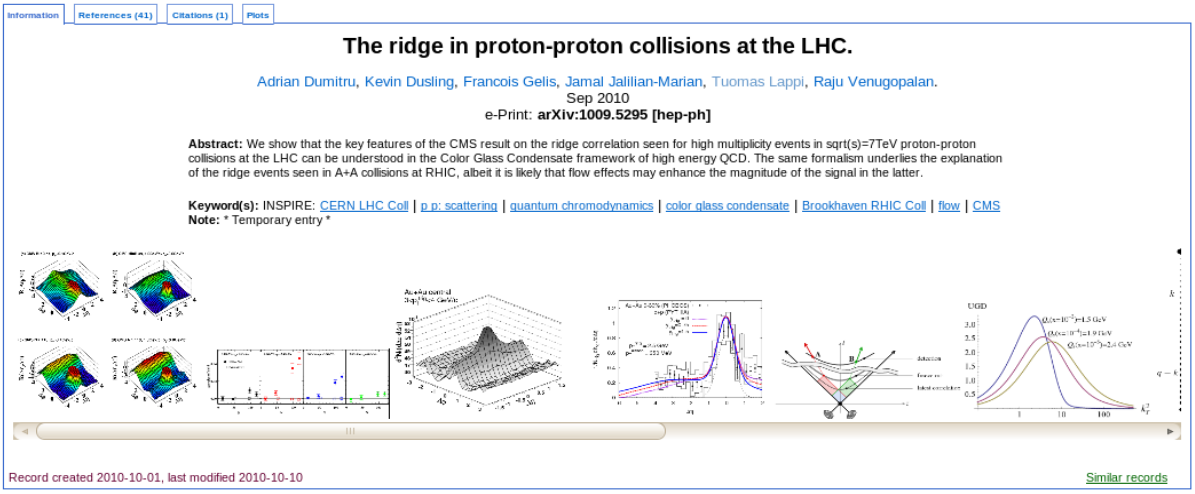

Figure 2: Article information with figures

INSPIRE database provides a comprehensive collection of relevant HEP documents and therefore citation analysis is one of the strong points in INSPIRE. For each article, INSPIRE shows the list of other articles that cite it. INSPIRE also has “Co-cited with” information which shows papers that are

frequently cited together. This enables users to easily find other related articles. Furthermore, a citation history graph shows citation counts of an article as a function of time. A citation information page is shown in figure 3.

Figure 3: INSPIRE citation information

INSPIRE author pages (example in figure 4) aim at providing comprehensive profiles about scientists in HEP. The information in author pages includes a breakdown of articles according to their type (published, review etc.), affiliation history, frequent keywords used and frequent co-authors. The lower part of an author page shows a breakdown of articles according to their citation counts. Hirsch index (h-index) [4] is calculated for each author as citation metric. Additional bibliograhic metrics will be added in order to measure the scientific importance of articles, authors, institutes and countries. Computationally intensive calculations of bibliometrics are executed in batch mode, using grid or cloud computing.

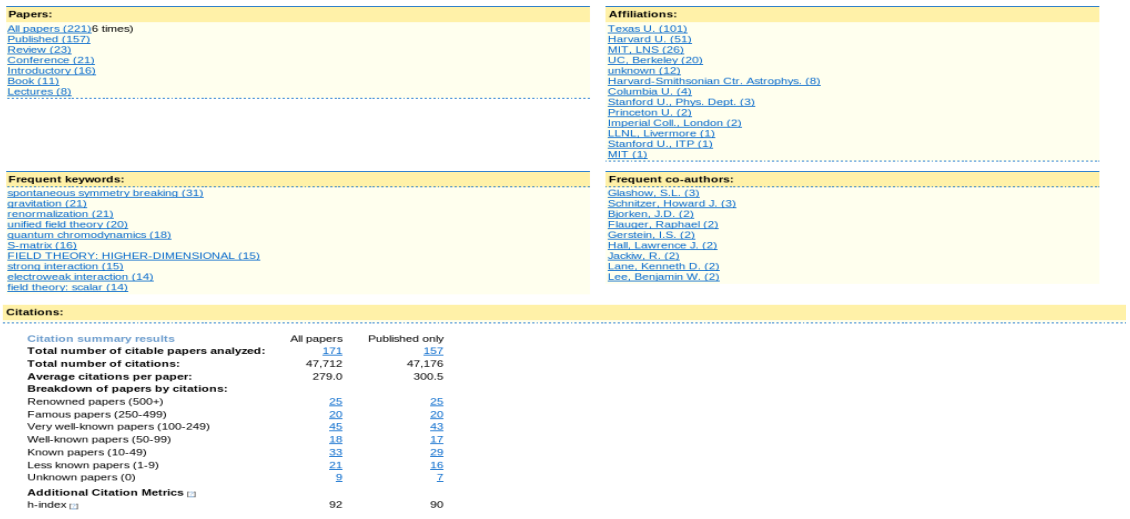

Figure 4: The INSPIRE author page

2. Invenio software

Invenio [5] is an integrated digital library system available under GNU General Public License. It consists of a set of modules for operating medium or large scale digital library services. Invenio technology covers all aspects of digital library management from document ingestion through to classification, indexing, curation and dissemination. Invenio is used to operate the CERN Document Server [6] with more than 500 collections and one million records covering articles, books, theses, journals, photos, videos etc. Invenio software is co-developed by an international collaboration comprising institutes such as CERN, DESY, EPFL, FNAL and SLAC and is used by about thirty scientific institutions worldwide. The software is based on a modular architecture and uses common standards like MARCXML [7] for storing bibliographic data and OAI-PMH for metadata exchange[8].

3. D4Science-II project

D4Science-II [9] is an EU funded project that develops and operates an e-Infrastructure supporting Virtual Research Environments (VREs). It develops an enabling technology (gCube [10]) which implements a mechanism for facilitating the interoperation of its e-Infrastructure with other autonomously running data e-Infrastructures. As a result, this creates the core of an e-Infrastructure ecosystem (figure 5). INSPIRE is one of the e-Infrastructures participating in D4Science-II project. In the context of the D4Science-II project, the INSPIRE e-Infrastructure makes available some of its resources and services to other members of the resulting ecosystem. Moreover, it benefits from the ecosystem via a dedicated Virtual Organization (VO) giving access to an array of resources ranging from computing and storage resources of grid infrastructures to data and services.

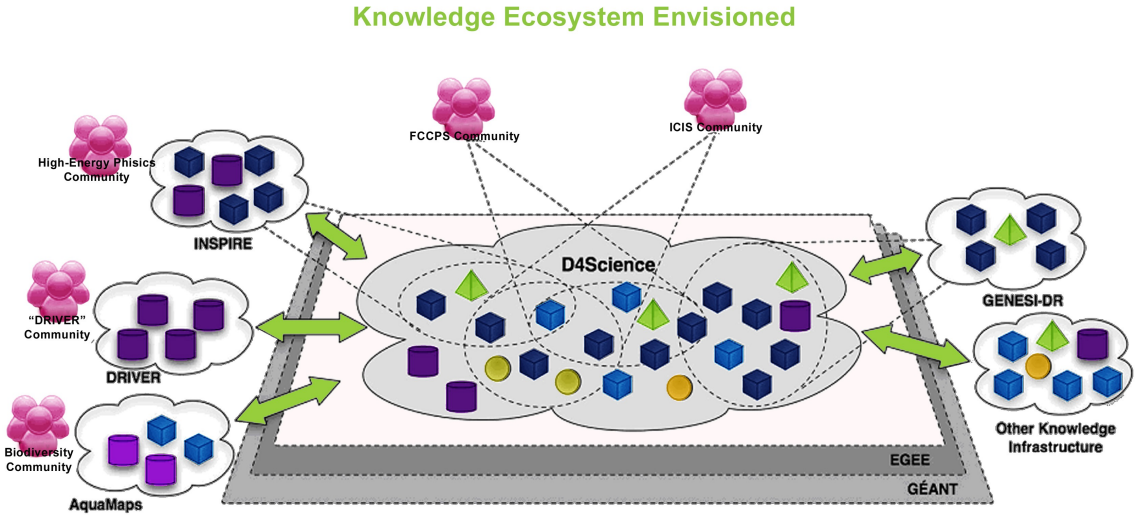

Figure 5: D4Science Knowledge Ecosystem

The D4Science infrastructure provides services based on the service oriented paradigm that can help in offering new functionality for INSPIRE users. D4Science Process Execution Engine (PE2ng) is a system which manages the execution of software in a distributed infrastructure. INSPIRE uses the different PE2ng adaptors to process INSPIRE data. PE2ng aims at bridging the gaps among different e-infrastructures and their users by allowing scientists to exploit computational resources via higher level constructs in a uniform manner. The adaptors provided by PE2ng include the JDLAdaptor for jobs inside D4Science infrastructure and the GridAdaptor for jobs in for example EGEE/EGI grid

environments. In addition, there are adaptors for Hadoop and Condor jobs. INSPIRE use cases in D4Science-II have used the JDLAdaptor, GridAdaptor and HadoopAdaptor for job management.

4. INSPIRE grid applications

Digital repositories like INSPIRE can profit from added computational power in several ways. Two major INSPIRE use cases in D4Science-II project are explained in this section. Other tasks that are executed on grid facilities include for example bibliometrics calculation and reference extraction.

4.1. Document OCR

Optical Character Recognition (OCR) is the translation of scanned documents into machine-encoded text. The CERN library and many other digital repositories have large numbers of scanned documents where textual information is not available. Therefore it is not possible to search for words or phrases in these documents, and applying techniques such as text mining is not possible.

OCR processes have often been handled using commercial services and tools, but now there are powerful open source tools for OCR. Using these tools the OCR process can be carried out in one workstation or by dividing the work in many parallel grid computing jobs. The OCR tool used in Invenio and D4Science-II is OCRopus [11]. OCRopus is a document analysis and OCR system, featuring pluggable layout analysis, pluggable character recognition, statistical natural language modeling, and multi-lingual capabilities. It is released under the Apache License and has a modular design through the use of plugins. OCRopus is also well suited for large scale batch processing so that the OCR tasks can be divided into independent grid jobs. A typical OCR process consists of selecting a set of scanned documents in pdf format, performing document layout analysis, line recognition and character identification. The output is in hOCR format (HTML document) which can be converted into pdf format.

An automatic procedure for the OCR process and grid job submission has been developed in Python and these tools are integrated as part of Invenio software. All the tools needed for OCR are available in one package (tar.gz file) that can be sent with the grid job or pre-installed in the grid nodes where jobs are executed. OCR jobs have been executed in local Linux workstations, the CERN lxbatch system, D4Science grid nodes and in EGEE/EGI grid infrastructures. Managing large numbers of parallel grid jobs can be a challenge and therefore a gridsubmit package [12] has been developed which allows for an easy definition of input files, submission to different grid backends and management of the grid jobs. Gridsubmit has been designed for problems like handling the OCR process of large document sets but it can also be used for other types of grid jobs.

4.2. Full-text indexing

The possibility to make a full-text search on the entire corpus of HEP literature is a very desired feature which was confirmed by a survey of the community [13]. In order to provide this functionality, INSPIRE needs to keep a mirror of the full text documents and use powerful indexing software as well as OCR for the old, scanned documents. Invenio's internal indexing tool is working in sequential mode. For the entire INSPIRE content to be indexed, it would take several months. The indexes are stored directly in the Invenio database meaning it would be very hard to compute them externally.

Motivated by the need and possibilities within the D4Science infrastructure, we decided to develop a prototype of parallel indexing using the Lucene library and the MapReduce model. Lucene [14] is a popular indexing solution broadly used in the library world. It is also well suited for parallelization and remote index calculation thanks to a portable format of index. The parallel processing framework is Hadoop [15]. It is an open source implementation of MapReduce [16] - the parallelization paradigm used by some of the biggest Internet companies which process large amounts of data. Use of Hadoop allows us to develop a very scalable solution which can be reused later for

other types of indexing or data processing. Job management effort is minimized in the sense that all job splitting, load balancing and resubmission are done by the Hadoop framework itself.

The currently developed prototype computes the Lucene full-text index of INSPIRE content using the Process Execution Engine's Hadoop adapter. The index can be later used in the INSPIRE server. Integration of the Lucene based full-text search in INSPIRE requires that for each mixed query (meta data and full text), the results need to be integrated. In practice it means that the full list of matching documents needs to be passed from Lucene server to Invenio in a fraction of a second. A low level solution is being tested to overcome this performance bottleneck.

5. Summary and Outlook

The beta version of INSPIRE is operational and provides the functionality of SPIRES as well as some new features. Processing power offered by grid and cloud computing help in offering new and better services for INSPIRE users. OCR and full-text indexing are the main use cases and are well suited for grid processing. Future use cases include reference extraction and calculating new bibliograhic metrics.

6. References

[1] http://www.slac.stanford.edu/spires/
[2] http://inspirebeta.net/
[3] http://www.projecthepinspire.net/
[4] Hirsch J. E. 2005 An index to quantify an individual's scientific research output PNAS 102 (46): 16569–16572.
[5] http://invenio-software.org/
[6] http://cdsweb.cern.ch/
[7] http://www.loc.gov/standards/marcxml
[8] http://www.openarchives.org/OAI/openarchivesprotocol.html
[9] http://www.d4science.eu/
[10] http://www.gcube-system.org/
[11] http://code.google.com/p/ocropus/
[12] Sompolski J 2010 The gridsubmit package & using grid for OCRing of documents, CERN-IT-Note-2010-001.
[13] Gentil-Beccot A et al. 2009 *J. Am. Soc. Inf. Sci. Technol.* **60** (2009) 150, arXiv:0804.2701
[14] http://lucene.apache.org/
[15] http://hadoop.apache.org/
[16] Dean J and Ghemawat S 2008 MapReduce: Simplified data processing on large clusters *Communications of the ACM* **51** 107—113

Acknowledgements

Authors would like to thank all the colleagues in INSPIRE and D4Science-II projects who have contributed to this work. Juliusz Sompolski and Christopher Hayward have worked as students and have provided valuable contributions.